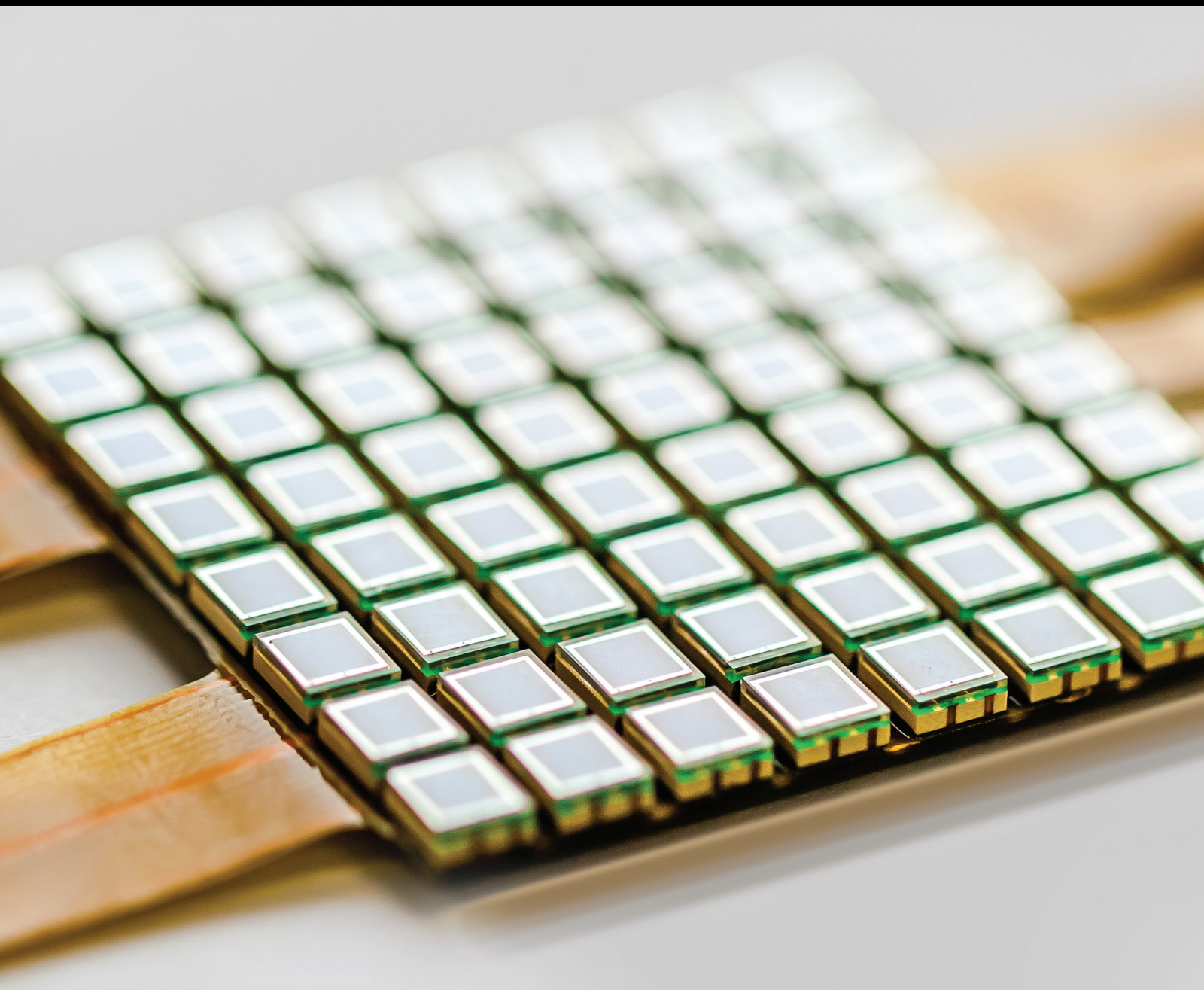
Journal of Sensors

Emerging Technologies: IoT, Big Data, and CPS with Sensory Systems

Lead Guest Editor: Qing Tan
Guest Editors: Nashwa El-Bendary, Magdy A. Bayoumi, Xiaokun Zhang, Javier Sedano, and José R. Villar



Emerging Technologies: IoT, Big Data, and CPS with Sensory Systems

Journal of Sensors

Emerging Technologies: IoT, Big Data, and CPS with Sensory Systems

Lead Guest Editor: Qing Tan
Guest Editors: Nashwa El-Bendary, Magdy A. Bayoumi, Xiaokun Zhang, Javier Sedano, and José R. Villar

Editorial Board

Contents

Hindawi
Journal of Sensors
Volume 2018, Article ID 3407542, 3 pages
https://doi.org/10.1155/2018/3407542

Editorial

Emerging Technologies: IoT, Big Data, and CPS with Sensory Systems

Qing Tan[1], **Nashwa El-Bendary**[2], **Magdy A. Bayoumi**[3], **Xiaokun Zhang**[1], **Javier Sedano**[4] and **José R. Villar**[5]

[1]*School of Computing and Information Systems, Athabasca University, Athabasca, AB, Canada*
[2]*Arab Academy for Science, Technology, and Maritime Transport, Smart Village, Giza, Egypt*
[3]*University of Louisiana at Lafayette, Lafayette, LA, USA*
[4]*Instituto Tecnológico de Castilla y León, Polg. Ind. Villalonquejar, López Bravo 70, 09001 Burgos, Spain*
[5]*Computer Science Department, University of Oviedo, EIMEM, c/ Independencia 13, 33004 Oviedo, Spain*

Correspondence should be addressed to Qing Tan; qingt@athabascau.ca

Received 5 December 2017; Accepted 5 December 2017; Published 18 February 2018

This collection of studies is focused on a three-legged stand in which now is focused on the research community: the internet of things (IoT), the cyber-physical systems (CPS), and the data-driven knowledge extraction based on big data. The availability of uniquely addressable heterogeneous electronic (UAHE)—including sensors, actuators, smart devices, RFID tags, embedded computers, and mobile devices—is continuously growing day by day. From a networking perspective, the IoT relies on interconnected UAHE for creating a mesh of devices, producing information, and building a worldwide network of real physical objects. In this context, the IoT presents a technology that enables loosely coupled decentralized systems of cooperating smart objects of autonomous physical-digital devices, augmented with sensing/actuating, processing, and networking capabilities.

Over the past 17 years, since the term IoT was first coined by Kevin Ashton in 1999, radio frequency identification (RFID) system is shaping up to be an important building block and a prerequisite for the IoT. The RFID is a key technology that automatically utilizes devices to identify, track and monitor, or control objects through radio waves. However, the performance of RFID equipment can be easily affected by factors in the surrounding environment. Consequently, that can impose a negative impact on the RF signal and the source data. H. Xu et al. proposed a sliding window cleaning algorithm called VSMURF, which is based on the traditional SMURF algorithm. In this research, the algorithm they proposed combines the dynamic change of RFID tags and the value analysis of confidence (δ) in order to solve the false negative problems affecting the quality of the collected data. Also, they used only one RFID reader and 25 tags for their experimental setting. The obtained experimental results showed that their proposed VSMURF algorithm outperformed the typical SMURF algorithm in most conditions, and when the tag's speed is low or high.

However, the frontiers between IoT and CPS are not always well defined. Some real world problems, like those related with eHealth and health monitoring, are in between: on the one hand, they gather information from the user environment with the aim of delivering the advice to either the user, the relatives, or the medical staff. P. M. Vergara et al. analyzed this problem in a real context of the epilepsy seizure detection. This research studies the problem of a system that, using wearable devices and Smartphones, copes with the detection of tonic-clonic seizures. The main factors in the design issues for such a platform are detailed. Finally, a description of a platform that is being developed is detailed, providing an experimentation on the robustness of the data communications. This experimentation would eventually

help in the online task distribution among the different computational resources.

Another example of the difficulty in establishing the limits of IoT and CPS is the problem of mobile crowd sensing: that is, using the sensory systems carried by each individual in a population in order to extract some relevant conclusion on a specific topic, like environment monitoring and positioning. No need to argue, this big data challenge as well as the problem of choosing the features, and how to perform the data fusion, and so on are topics covered with big data techniques. Nevertheless, L.-Y. Jiang et al. studied the problem of the trust ability of the data as a problem of avoiding individuals to introduce false data—either intentionally or unintentionally. Furthermore, the authors propose an incentive system to tackle this issue. The proposed incentive system is based on the conditions of recruiters and candidates, becoming in a quality assessment optimization problem solved by a polynomial-time greedy approximation algorithm the authors proposed. The simulation-based experimentation is complete and full of interest.

One of the problems that the IoT research community needs to tackle is to ensure the security: not only the problems concerning the unauthorized and/or destructive actions on the devices and/or infrastructure but also those related with cyberattacks and any kind of malicious strikes. Surprisingly, this issue has not been studied in depth due to the variability of the IoT scenarios: different communication hardware and software, the computation restrictions in the nodes, and so on. This issue has been analyzed in T. Andrysiak et al., proposing several interesting concepts that need to be considered when dealing with security issues. Furthermore, the authors proposed a very interesting method for detecting anomalies in the network traffic that might arise from cyberattacks. To do so, the authors proposed to estimate the normal network behaviour by (i) filtering outliers; (ii) computing the exponential smoothing models, either with Brown's, Holt's, or Winter's models; (iii) analysing the fluctuations of the estimations using Bollinger's bands; and (iv) comparing the obtained estimations of normal behaviour with what is really happening in the network. A very complete experimentation is performed using smart metering networks, a real IoT problem indeed.

Besides, challenges in the deployment of IoT solutions are also focused on energy efficiency. In passive backscatter communication systems, sensor nodes need to harvest RF energy from transceivers or readers and use it to recharge their finite energy storage capacity. However, RF energy harvesting also faces unpredictable environmental challenges, which make the sensors' sensing, processing, and communication activities difficult. Therefore, an efficient energy management is necessary in order to guarantee the sensors' activities and QoS (quality of service) of backscatter communication. In this special issue, the article contributed by S. Hu et al. present an RF energy harvesting and a colocated passive quadrature amplitude modulation (QAM) backscatter communication signal models. The proposed models aim at achieving good quality of service (QoS) level through minimizing interference resulted from multisensor and increasing spectral efficiency. Experimental results, based on numerical analysis and simulations, showed that the maximum throughput inversely relates to the consumed power and the number of sensors. Also, for a given consumed power of sensors, it was observed that the throughput decreases with duty cycle, and the number of sensors has little effect on the throughput.

On the other hand, the coupling between the IoT and the big data communities is strong as big data analytics has become an essential component for extracting value from data. Big data as one of the most important and recent research challenges with a paradigm relies on the collection of tremendous amount of data to support innovation in the upcoming decades. A dataset is considered as big data when it meets the "four Vs" requirements, namely, volume, variety, velocity, and value. The keystone of big data exploitation is to leverage the existing datasets to create new information and predict future happenings, enriching the decision value chain. Accordingly, as the IoT continuously collects data about the surrounding living environments, it is considered as a prototypical example of big data and a great application area of big data analytics.

Dealing with large amount of dynamic data is one of important characteristics of big data analytics. Entity resolution (ER) is to disambiguate manifestations of real-world entities in the dataset by linking and grouping and to reduce the complexity of data processing, which is a crucial step to achieve effective, efficient, and accurate data processing in big data analysis. H.-J. Zhu et al. proposed a type-based multiblock technique to improve data quality. In the article, they present a new ER solution using a hybrid approach. In particular, through their new ER workflow, it is capable to reduce the searching space for entity pairs. The performance of their proposed method has been validated through using a real-life dataset created from an IoT real project. These results have been obtained using up to five standard metrics. The experimental results conclude that their proposed approach can be a promising alternative for ER over large-scale data.

Besides, CPS are emerging from the integration of embedded computing devices, smart objects, people, and physical environments, which are typically tied by a communication infrastructure. So, the design of CPS and the implementation of their applications need to rely on IoT-enabled architectures and protocols that, both locally and globally, enable collecting, managing, and processing large data sets and support complex processes to manage and control such systems. Thus, as a matter of fact, the large-scale nature of IoT-based CPS can be effectively and efficiently facilitated and supported via utilizing the cloud computing infrastructures and platforms for providing flexible computational power, resource virtualization, and high-capacity storage for data streams in addition to ensuring safety, security, and privacy.

CPS are the extension of traditional closed-loop systems with the computer-based control and the Internet-based communication, in which physical components, computer algorithms, and sensing elements are tightly integrated via the Internet communication link. Sensors are the essential

elements in any closed-loop system. In this special issue, the article contributed by L. Sanchez et al. present a model-based virtual sensor to be used for Li-Ion batteries' condition monitoring in cyber-physical vehicle systems. A principle-based model encoded with the expert knowledge about battery behaviour has been turned into a soft sensor. Feeding with the information extracted from data collected from on-vehicle measurements, the soft sensor is able to approximate the state of health of a battery. The experimental results coming from the implementation of the model-based soft sensor for fault detection and diagnosis of the batteries have demonstrated the soft sensor's high efficiency.

As an example of CPS, W. Yu et al. proposed a new approach to generate flexible gait for bionic leg's control based on muscle synergies extracted from sEMG (surface electromyogram) signal. The approach was inspired by the fact that muscle synergies leading to dimensionality reduction may simplify motor control and learning. Usually, typical multichannel sEMG signal-based controls need mass data and vary greatly with time which cause worse latency and other performance-related issues and make it difficult to generate compliant gait. The paper addresses two questions to highlight the essential features in the proposed approach with an interesting experimental result. The first question addresses whether the same set of muscle synergies can explain the different phases of jumping movement with various velocities. The second question is about building a model for generating velocity-adapted jumping gait with muscle synergies, in which a wavelet neural network is proposed to predict the reference gait pattern, while fuzzy inference system is adopted to merge these reference gaits in order to create more generalized gaits with different jumping rhythms. The proposed method can be adopted as the decoder in sEMG-based controls for a bionic leg. Moreover, linear combinations of synergies may describe complex force and motion patterns in reduced dimensions, and the robust representations of synergies within the control scheme can generate flexible gaits for other complex motions.

Qing Tan
Nashwa El-Bendary
Magdy A. Bayoumi
Xiaokun Zhang
Javier Sedano
José R. Villar

Hindawi
Journal of Sensors
Volume 2018, Article ID 2094696, 12 pages
https://doi.org/10.1155/2018/2094696

Research Article

A Type-Based Blocking Technique for Efficient Entity Resolution over Large-Scale Data

Hui-Juan Zhu,[1] **Zheng-Wei Zhu**,[1] **Tong-Hai Jiang**,[2,3] **Li Cheng**,[2,3] **Wei-Lei Shi**,[2] **Xi Zhou**,[2,3] **Fan Zhao**,[2,3] **and Bo Ma**[2,3]

[1]*School of Information Science & Engineering, Changzhou University, Changzhou 213164, China*
[2]*The Xinjiang Technical Institute of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China*
[3]*Xinjiang Laboratory of Minority Speech and Language Information Processing, Urumqi 830011, China*

Correspondence should be addressed to Zheng-Wei Zhu; zhuzw@cczu.edu.cn and Tong-Hai Jiang; jth@ms.xjb.ac.cn

Received 21 April 2017; Accepted 11 September 2017; Published 28 January 2018

Academic Editor: Qing Tan

In data integration, entity resolution is an important technique to improve data quality. Existing researches typically assume that the target dataset only contain string-type data and use single similarity metric. For larger high-dimensional dataset, redundant information needs to be verified using traditional blocking or windowing techniques. In this work, we propose a novel ER-resolving method using a hybrid approach, including type-based multiblocks, varying window size, and more flexible similarity metrics. In our new ER workflow, we reduce the searching space for entity pairs by the constraint of redundant attributes and matching likelihood. We develop a reference implementation of our proposed approach and validate its performance using real-life dataset from one Internet of Things project. We evaluate the data processing system using five standard metrics including effectiveness, efficiency, accuracy, recall, and precision. Experimental results indicate that the proposed approach could be a promising alternative for entity resolution and could be feasibly applied in real-world data cleaning for large datasets.

1. Introduction

Entity resolution (ER), also known as record linkage, entity reconciliation, or merge/purge, is the procedure of identifying a group of entities (records) representing the same real-world entity [1–3]. Generally speaking, ER has become the first step of data processing and widely used in many application domain, such as digital libraries, smart city, financial transactions, and social networks. Especially with the rapid development of Internet of Things technology, it is common that data contains a huge amount of inaccurate information and different types of ambiguities [4]. Accordingly, developing proper ER techniques to clear and integrate data collected from multiple sources has received much attention [1, 5, 6]. The ultimate goal of ER technologies is to improve data quality or to enrich data to facilitate more detailed data analysis.

Researchers have proposed many automatic methods and techniques to resolve ER problem across multiple resources, to detect if they refer to the same entity and therefore can be merged. Wang and Madnick [7] proposed a rule-based method by using rules and unique key attribute developed by experts. However, this kind of method has some additional restrictions, such as the result of the rules must always be correct. Bilenko and Mooney [8] proposed an adaptive duplicate detection method, but it mainly focus on string similarity measures. Recently, based on different similarity metrics, some researchers proposed machine learning methods to classify the entity pairs to "match," "non-match," or "possible-match", such as the [9] proposed automatic record linkage tool by using of support vector machine classification of Christen. However, this kind of methods requires a large amount of manually labelled data, and if the entity pairs are classified to "possible-match," a manual review process is required to assess and classify them into "match" or "non-match." Thus, this is usually a time-consuming, cumbersome, and error-prone process [10].

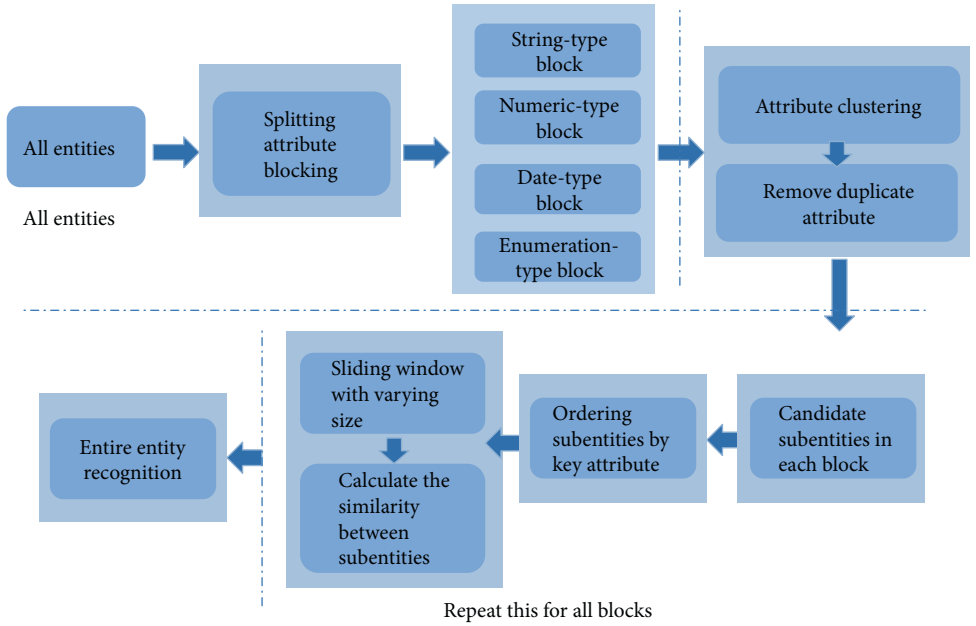


Figure 1: The Overall framework of the proposed entity-resolution technique.

The mentioned work in ER mainly focused on the development of automatic algorithms [11]. A general procedure for solving the ER problem includes calculating the similarity of all entity pairs by means of similarity metric methods [12], such as Jaccard similarity coefficient and Levenshtein distance. Entities whose similarity values are higher than the specified threshold are considered to be the same entity. However, in the face of large datasets, the performance deteriorates drastically as a comparison of dense attributes [2]. To solve this problem, blocking and windowing mechanisms [2, 10, 13, 14] have been introduced. The goal of the blocking scheme is to group entities into block-based clustering technique to reduce comparison searching space into one block. Windowing methods, such as sorted neighborhood method (SNM) [13, 15] or multipass sorted neighborhood method (MPN) [16, 17], sort entities according to the keywords, and then slide a fixed size window over them to compare the entities within the window.

Algorithmic approaches have been improving in quality, but there are still some problems that have not been fully studied [1]. First, existing research typically assumes that the target dataset only contain string-type data and use single similarity metric [1, 11, 18, 19]. Second, for large high-dimensional target dataset, redundant information is verified, which not only increases the computational complexity but also may deteriorate the quality of ER results [20]. Third, the common drawback of most blocking techniques is that they put the complete entities in one block (or multiple blocks) according to the block keywords, which leads to the lack of the necessary flexibility for composing ER workflows of higher performance in combination with specialized, complementary methods [5, 21]. Fourth, the performance of typical windowing methods (SNM or MPN) depends strongly on the size of sliding window [21], but they often employ a fixed window size. The larger the fixed window size, the more comparisons are executed, and the lower the overall efficiency gets, however, small size may lead to a high number of missed matches (e.g., the closest entities are not placed in the same window) and to low effectiveness [2, 21–23].

Based on above problems, this paper introduces a novel ER resolving method. First, we introduce a novel blocking scheme based on attribute value types. In these different type of blocks, we adopt diverse similarity metrics to achieve maximum efficiency gains, because even though a metric method showed robust and high performance for one type of data [24, 25], it may perform poorly on others [26–28]. Second, we provide an attribute clustering method, a novel and effective blocking approach, which is a preprocessing scheme to resolve ER in a significantly lower redundancy and higher efficiency. Its core is to divide attribute names into nonoverlapping clusters according to the similarity or correlation of the attribute values. Third, we introduce a comparison mechanism, which combines a dynamically adjustable window size to specify the processing order of all individual comparisons in each block. It aims at identifying the closest pairs of entities involving significantly fewer comparisons while maintaining the original level of effectiveness. Fourth, during the ER completion step, we adopt a weighted undirected graph to gather the output of each block and used the improved weighting and pruning scheme to enhance its efficiency. Finally, we assess the performance of our methods with using a real-life dataset, and the experimental results validate the exceptional performance of our methods.

2. The Overall Framework

The overall framework of the proposed ER resolving method is shown in Figure 1. It consists of three parts: the first part

divides the complete entities into different blocks according to the attribute value types; the second part introduces the method of attribute clustering, which is used to remove redundant attributes; the third part performs comparisons solely between the subentities within the scalable window in each block and gathers the outputs from individual blocks to finish the ER problem.

3. Type-Based Blocking Approach

At the core of our approach lies in the notion of type-based blocking (defined as Definition 1), varying window size and blocking graph. Given an entity set R which is defined as definition 2, the common attribute value types include string, numerical, enumeration, and date. In order to avoid the defects of fixed window size, the proposed method uses different dynamic scalable window strategies (refer to Section 3.3.2) in different type of blocks, also known as varying window size. In addition, given a Map *closestMap*, which is used to store the closest entities, the corresponding blocking graph G_e (also known as edge-weight graph) is constructed according to the value of *closestMap*: each entity identifier contained in *closestMap* is mapped to a node in the blocking graph, and each pair of cooccurring entities (e.g., entities that are marked as matched at least in one block) is connected with an undirected edge.

Definition 1. (Type-Based Blocking) In this paper, according to attribute value types, the dataset is divided into the following blocks: numeric-type block, string-type block, date-type block, and enumeration-type block, and these blocks are nonoverlapping.

Definition 2. (Entity Set) Entity set is defined as $R = \{r\_1, r\_2, \ldots, r\_n\}$, and each entity can be expressed as $r\_i = \{x\_i1, x\_i2, \ldots, x\_im\}$ $(1 \le i \le n)$. The attribute vector is defined as $A = \{A\_1, A\_2, \ldots, A\_m\}$, and A_l $(1 \le l \le m)$ is used to represent the lth attribute. Accordingly, the value of A_l in entity r_i is represented as x_il $(1 \le l \le m,\ 1 \le i \le n)$.

Definition 3. (Blocking Graph) Given an entity collection *closestMap*, the undirected blocking graph $G\_e = \{V\_e, E\_e, \mathrm{WS}\}$ is generated by it, where V_e is the set of its nodes, E_e is the set of its undirected edges, and WS is the weighting scheme that determines the weight of every edge.

3.1. Splitting Attributes to Different Blocks. The comparison between attribute values is an important task for resolving the ER problem. A variety of methods has been developed for that [29–32], which typically rely on string comparison techniques. However, the dataset might contain other types of data, such as numerical, enumeration, and date, and there is still much work to be done about these types of data [33–35].

The purpose of splitting attributes into different blocks is to facilitate the introduction of a variety of flexible similarity metrics to identify redundant attributes or closest entities. For instance, entities (Jimi, F) and (Jimi, M), in which "F" and "M," respectively, represent the "female" and "male" for the gender attribute. The similarity between these two entities achieved by traditional methods, such as Levenshtein distance [36], is very high. However, these two entities represent two individuals with same name but with different gender. The attribute gender is an enumeration type, and the similarity of the enumeration type should be calculated using equal or unequal, rather than the methods often used for string-type data. The functionality of splitting attributes into different blocks is outlined in Algorithm 1.

3.2. Attribute Clustering Method. The increase of data dimensionality brings new challenges to the efficiency and effectiveness of many existing ER methods. In this article, we introduce an attribute clustering method to remove redundant attributes. As a result, we obtain a more compact and easily interpretable representation of the target concept. The core of the proposed attribute clustering method is to divide attributes into disjoint clusters according to the similarity or correlation of the attribute values within the same type block, which is based on the concept of redundant. This scheme offers a better balance between the computational cost and the precision for resolving ER [37]. In order to facilitate the discussion, the following definitions are introduced.

3.2.1. String-Type Block. The attribute clustering method for string type depends on two parts: (1) the model that congruously represents the values of attribute; (2) the similarity metric that catches the common pattern between the two sets of attribute values. In this paper, the weight of each term is obtained by TF-IDF (term frequency-inverse document frequency) algorithm [38]. Accordingly, the two sets of attribute values can be represented as $v\_i = \{wi\_1, wi\_2, \ldots, wi\_n\}$ and $v\_j = \{wj\_1, wj\_2, \ldots, wj\_n\}$, and θ is the angle between v_i and v_j. Thus, the similarity between two sets of attribute values with cosine similarity can be defined as

$$\mathrm{sim}(v\_i, v\_j) = \cos\theta = \frac{\sum_{t=1}^{n} (wi\_t.wj\_t)}{\sqrt{\sum_{t=1}^{n} wi\_t^2 . \sum_{t=1}^{n} wj\_t^2}}, \quad (1)$$

where $\mathrm{sim}(v\_i, v\_j)$ takes values in the interval [0, 1], with higher values indicating higher similarity between the given attributes. If it is higher than an established threshold, that means the given attributes are redundant.

3.2.2. Numeric-Type Block. In this paper, the numerical attributes are divided into three categories [34]: (1) nominal attribute, it refers to only a set of numerical symbols, such as ID card number, which does not have a meaningful order, and is not quantitative; (2) ordinal attribute, its possible values have the meaningful orders or ranking, but the difference between successive is unknown; (3) general numerical attribute, its possible values are measurable, such as age and height. Assume that there are two numerical attributes A and B, and the values of A is expressed as $\{a\_1, a\_2, \ldots, a\_c\}$ and the values of B is expressed as $\{b\_1, b\_2, \ldots, b\_r\}$. The correlation between

```
Input: sd: dataset;
       δ: the threshold for the number of possible values of an enumeration attribute;
       λ: the number of values which are randomly selected from sd;
Output: Map < BT(block type), list of attribute names > BT∈{NUME, STRING, DATE, ENUM}
(a)  Map < attribute, List < v1, v2,…, vλ> > mediateData ← sd;// Using Map to store attribute and its
     values.
(b)  blockMap← new HashMap < String, List>; // blockMap is used to store the return value;
(c)  For each attribute in mediateData Do
(d)       valuesNoRep← Remove duplicate elements from List < v1,v2,…,vλ>;
(e)       n← valuesNoRep.size();
(f)       If ((double)n/λ < δ) then {//the type of this attribute is enumeration
(g)            List enumAttributes ← blockMap.get("ENUM");
(h)            If (enumAttributes == null) then {enumAttributes ← new ArrayList;
(i)                 blockMap.put("ENUM", enumAttributes);}
(j)            enumAttributes← enumAttributes.add(attribute name);
(k)       }Else if (the elements of valuesNoRep conform to the date type rules) then {
(l)            List dateAttributes ← blockMap.get("DATE");
(m)            If (dateAttributes == null) then {dateAttributes ← new ArrayList;
(n)                blockMap.put("DATE", dateAttributes);}
(o)            dateAttributes ← dateAttributes.add(attribute name);
(p)       }Else if (The elements of listWithoutDu conform to the numerical type rules) then {
(q)            List numericAttributes ← blockMap.get("NUME");
(r)            If (numericAttributes == null) then {numericAttributes ← new ArrayList;
(s))               blockMap.put("NUME", numericAttributes);}
(t)            numericAttributes ← numericAttributes. add(attribute name);
(u)       } Else {
(v)            //other attributes will be treated as string type
(w)            List stringAttributes ← blockMap.get("STRING");
(x)            If (stringAttributes == null) then {stringAttributes ← new ArrayList;
(y)                blockMap.put("STRING", stringAttributes);}
(z)            stringAttributes ← stringAttributes.add(attribute name);
(aa)      }
(bb) End For each attribute in mediateData
(cc) return blockMap;
(Note: the δ and λ should be adjusted according to the size of dataset)
```

Algorithm 1: Splitting attributes into different blocks.

two nominal numerical attributes is discovered by χ^2 test [39], which is shown in

$$\chi^2 = \sum_{i=1}^{c}\sum_{j=1}^{r}\frac{(o\_ij - e\_ij)^2}{e\_ij}, \tag{2}$$

$$e\_ij = \frac{\text{count}(A = a\_i) \times \text{count}(B = b\_j)}{n}, \tag{3}$$

where o_ij is the observation frequency and e_ij is the expected frequency. The calculation method of e_ij can be referred to (3), where n is the number of data tuple and $\text{count}(A = a\_i)$ is the count of value a_i which appeared in the A attribute.

For the general numerical attribute, their correlation is measured by Pearson correlation coefficient [40], which is shown in

$$r\_A, B = \frac{\sum_{i=1}^{n}(a\_i - \bar{A})(b_i - \bar{B})}{n \cdot \sigma\_A \cdot \sigma\_B} = \frac{\sum_{i=1}^{n}(a\_i \cdot b\_i) - n\overline{A}\cdot\overline{B}}{n \cdot \sigma\_A \cdot \sigma\_B}, \tag{4}$$

where σ_A and σ_B are the standard deviation of attributes A and B. $\bar{A}$ and $\bar{B}$ are the mean values. r_A, B is their correlation coefficient, whose value space is $[-1,1]$. If r_A, B is bigger than the established threshold, that means one of the given attributes can be removed as redundant [34].

For the ordinal numerical attribute, it is divided three steps: (1) we sort and divide the values of attribute A into set $\{M\_1, M\_2, \ldots, M\_100\}$, also known as percentile [41, 42]. The processing of the attribute B is the same as that of A; (2) we count the number of attribute values that fall into each interval and use this count value ca_i to replace M_i $(1 \le i \le 100)$. Thus, the replacement set $\{ca\_1, ca\_1, \ldots, ca\_100\}$ and $\{cb\_1, cb\_1, \ldots, cb\_100\}$ can be obtained; (3) $\{ca\_1, ca\_1, \ldots, ca\_100\}$ and $\{cb\_1, cb\_1, \ldots, cb\_100\}$ can be regarded as the general numerical attribute to calculate their correlation.

3.2.3. Date-Type Block. The correlation between date-type attributes, which is the handling process, is same with that of the ordinal numerical attribute.

Input: Map < attribute name, List $\langle v\_1, v\_2, \ldots, v\_n \rangle$ >//this is for same type block
δ: the threshold for the attributes similarity
Output: Set of attribute names clusters: C
(a) *connects* ←{};
(b) noOfAttributes ←the size of Map < attribute name, List $\langle v\_1, v\_2, \ldots, v\_n \rangle$ >;
(c) For (int i = 0; i < noOfAttributes; i++) do
(d) For (int j = i + 1; j < noOfAttributes; j++) do
(e) *Sim_i*, *j* ←attribute[i].getSimilarAttribute(attribute[j]);//refer to the formula (1)~(4)
(f) If (*Sim_i*, *j* > δ) then *connects* ← *connects*.add(new Connect(i, j));
(g) End For(j)
(h) End For (i)
(i) cons ←computerTransitiveClosure(*connects*);
(j) *C*←getConnectedComponents(cons);
(k) For each *c_i* ∈ *C* do
(l) If (*c_i*.size() == 1) then *C*.remove(*c_i*);
(m) End For (*c_i*);
(n) Return *C*;

Algorithm 2. Attribute clustering algorithm.

3.2.4. Enumeration-Type Block. For the numeration-type attributes, their correlation measurement is same with that of the nominal numerical attribute.

The detailed functionality of the proposed attribute clustering method is described in Algorithm 2, which is based on the method proposed by Papadakis et al. in research [43].

In principle, the attribute clustering method proposed in this paper works as follows: each attribute name in the input *map* is associated with the most similar or the strongest correlation attribute name (lines c–h). The *Connect* between two attribute names is stored in a data structure on the condition that the similarity or correlation of their values is more than the threshold δ (line f). The transitive closure of the stored *connects* is then obtained to build the basis for partitioning attribute names into cluster (line i). As a result, each cluster is taken from each connected component of the transitive closure (line j), and the singleton clusters are removed from *C* (lines k–m).

3.3. The Closest Entity Detection. SNM was first proposed in the mid-1990s [44]. Its basic idea is to sort the dataset (entity set) according to the sort keywords and to sequentially move a fixed size window over the sorted entities. Candidate entity pairs are then generated only from entities within the same window. This technique reduces the complexity from $O(n \times n)$ to $O(n \times w)$, where n is the number of input entities and w is the window size. The linear complexity makes SNM more robust against load balancing problems [16, 17]. However, this method still has a room for improvement, such as the following:

(a) When some attribute values are missing or the length difference between the attribute values is larger, the weighting summation method for calculating the similarity between entities will no longer appropriate.

(b) The interference caused by redundant attributes will increase the comparison's cost and affect and the quality of ER results [20].

(c) The comparison strategy in SNM is that when each new entity enters the current window, it needs to be compared with the previous w-1 (w is the window size) entities to find "matching" entities. However, it is difficult to determine the sliding window size [2].

(d) The real-life dataset often contains a variety of data types, such as enumeration, numerical, date, and string, but the traditional SNM usually assumes that the dataset only contains string-type data and just employ single similarity metric based on string type [11].

To solve the above problems, this paper proposes multiblocking sorted neighborhood (MBN) algorithm. In order to facilitate the discussion, the definition of valid weight is introduced.

Definition 4. (Valid Weight) When comparing the attribute values *r_il* and *r_jl*, if one of them is missing or their length ratio is lower than the specified threshold, the valid weight $\vartheta\_ij^{(l)}$ will be set to 0, the length ratio are shown in (5), and the length ratio is only adopted in string- and enumeration-type blocks:

$$\text{lenRatio} = \frac{\text{len\_big}}{\text{len\_small}}, \tag{5}$$

$$\text{len\_big} = \text{Max}(\text{length}(r\_il), \text{length}(r\_jl)), \tag{6}$$

$$\text{len\_small} = \text{Min}(\text{length}(r\_il), \text{length}(r\_jl)). \tag{7}$$

3.3.1. The Similarity Metrics. As mentioned above, the data types in this article include string, numeric, date, and enumeration. Attribute value similarity metrics are discussed based on these four types.

(1) String-Type Attribute. Cosine similarity [45–47] is used to calculate the similarity of attribute value *x_il* and *x_jl*, which can be defined as sim_att_string(x_il, x_jl). Accordingly, the similarity of entity *r_i* and *r_j* is

calculated by (8), and t is the number of attributes in this string-type block.

$$\text{sim\_ent\_str}(r\_i, r\_j) = \frac{\sum_{l=1}^{t} \vartheta\_ij^{(l)} \text{sim\_att\_str}(x\_il, x\_jl)}{\sum_{l=1}^{t} \vartheta\_ij^{(l)}}. \quad (8)$$

(2) Numeric-Type Attribute. The numeric type is divided into nominal, ordinal, and general numerical as mentioned above. The diversity is adopted in numeric type, and the diversity of attribute values x_il and x_jl from attribute A_l ($1 \le i, j \le n$, $1 \le l \le m$) is defined as div_ij^l. The calculation methods are defined as the following:

(a) A_l is one of the general numerical attributes, $\text{div}\_ij^l = |x\_il - x\_jl|/(\max\_A\_l - \min\_A\_l)$, where $\max\_A\_l$ is the maximum value of A_l and $\min\_A\_l$ is the minimum one.

(b) A_l is one of the ordinal numerical attributes, and A_l has the ordered state M which can be expressed as $\{1, \ldots, M\_k\}$. Using the corresponding ranking number s_il (when the $x\_il \in M\_r$ ($1 \le r \le k$), the ranking number of x_il is r) to replace x_il and calculating $z\_il = (s\_il - 1)/(M\_r - 1)$ accordingly, next, s_il is replaced by z_il. Finally, the processing of z_il can refer to step (a).

(c) A_l is one of the nominal attributes: if $x\_il = x\_jl$, then $\text{div}\_ij^l = 0$; else $\text{div}\_ij^l = 1$.

Consequently, the diversity of entity r_i and r_j is defined in (9), and the similarity between them is shown in (10), where r is the total number of attributes in this numeric block.

$$\text{div\_ent}(r\_i, r\_j) = \frac{\sum_{l=1}^{r} \vartheta\_ij^{(l)} \ \text{div}\_ij^{(l)}}{\sum_{l=1}^{r} \vartheta\_ij^{(l)}}, \quad (9)$$

$$\text{sim\_ent\_num}(r\_i, r\_j) = 1 - \text{div\_ent}(r\_i, r\_j). \quad (10)$$

(3) Enumeration-Type Attribute. Because the possible values for enumeration attribute are a set of predefined specific symbols, there is usually no implicit semantic relationship between them. This article uses equality or inequality to measure the similarity between two enumerated attribute values, which is shown as follows:

$$\text{sim\_att\_enum}(x\_il, x\_jl) = \begin{cases} 0 & (x\_il \ne x\_jl), \\ 1 & (x\_il = x\_jl). \end{cases} \quad (11)$$

Accordingly, the similarity of entity r_i and r_j is obtained by (12), where s is the amount of attributes in enumeration-type block.

$$\text{sim\_ent\_enum}(r\_i, r\_j) = \frac{\sum_{l=1}^{s} \vartheta\_ij^{(l)} \text{sim\_att\_enum}(x\_il, x\_jl)}{\sum_{l=1}^{s} \vartheta\_ij^{(l)}}. \quad (12)$$

(4) Date-Type Attribute. The processing of date-type attributes, firstly, is to divide the attribute values into a set of ordered states $\{1, \ldots, M\_k\}$. Correspondingly, the similarity between x_il and x_jl from the date attribute A_l is defined as (13), and the similarity of entity r_i and r_j is calculated by (14).

$$\text{sim\_att\_date}(x\_il, x\_jl) = \begin{cases} 1 & (x\_il \in M\_i, x\_jl \in M\_i), \\ 0 & (x\_il \in M\_i, x\_jl \in M\_j, i \ne j), \end{cases} \quad (13)$$

$$\text{sim\_ent\_date}(r\_i, r\_j) = \frac{\sum_{l=1}^{s} \vartheta\_ij^{(l)} \text{sim\_att\_date}(x\_il, x\_jl)}{\sum_{l=1}^{s} \vartheta\_ij^{(l)}}. \quad (14)$$

3.3.2. The Varying Window Sizing Strategies. The window size ws in traditional SNM is fixed, and the challenge here is how to select a reasonable value for ws. If the window size is too large that will increase the comparison cost, in contrast, too small may lead to missing matches. In this paper, depending on the type-based blocks, we use varied strategies to dynamically adjust the window size in these blocks.

For string-type block, we adopt the method proposed by [15] to dynamically adjust the window size based on the amount of closest entities in the current window, which can be defined as follows:

$$\text{ws\_}i = \text{Int}\left(\text{ws\_min} + \frac{n\_\text{actclo}}{\text{ws\_}(i\text{-}1) - 1}(\text{ws\_max} - \text{ws\_min})\right), \quad (15)$$

where ws_i is the window size of the ith window, ws_min is the minimum of window size and $W_{\max}$ is the maximum one, and n_actclo is the number of the closest entities. But the calculation method of the minimum and maximum window size is not presented in research [15]. In this paper, we obtained the minimum and maximum with the following steps: (1) selecting the key attribute; (2) clustering the values of the key attribute based on *Cosine* similarity metric; (3) ws_min is equal to the number of elements of the minimum cluster and ws_max is equal to the number of elements in the biggest cluster.

For numerical-type and date-type block, the conditions for expanding window size are $(1\text{-div}\_iw_i^{(l)}) > \delta$ and $\text{sim\_att\_date}(x\_il, x\_w_il) > \delta$, and the conditions for narrowing window size are $(1\text{-div}\_iw_p^{(l)}) < \delta$ and $\text{sim\_att\_date}(x\_il, x\_w_pl) < \delta$, where x_il is from the entity of the latest sliding into the current window, x_w_il is from the last one in the current window, and x_w_pl is from the pth entity in the current window. In addition, the expansion step ws_step (ws_step = 1) and the minimal window size ws_min are also set in advance. Accordingly, the expanded window size ws_i is shown in (16), and the narrowed window size is described by (17).

$$\text{ws\_}i = \text{ws\_}i + \text{ws\_step}, \quad (16)$$

Input: *db*: dataset in each block
*ws*_min: the minimal window size(the string type block need also provide *ws*_max)
*A*_*l*: the sort key attribute in each block
δ: the similarity threshold in each block
Output: Map < sub record identifier, Set < the identifiers of the closest entities>>
(a) *winSize*←*ws*_min;
(b) *type_att* ← the type of *A*_*l*;
(c) if (*type_att*.equals("STRING")) then *winSize* ←*ws*_max;
(d) else if (*type_att*.equals("ENUM")) then *winSize* ←calculate the window size according to (18);
(e) Sorting entities in each block based on *A*_*l*;
(f) Sliding the window from the first entity in each block;
(g) *closestMap*←new HashMap < Integer, Set < Integer>>;
(h) While (it is not the last entity) Do
(i) *i*←0; //initialize the comparison times
(j) *latestEnt*←the latest entity that sliding into the current window;
(k) While (*i* < *winSize*) Do
//according to the (8), (10), (12), and (14) to calculate the similarity between entities
(l) *similarity*←the similarity between entity *latestEnt* and the i^{th} entity in current window;
(m) If (*similarity* > δ) then Do{
(n) *closestSet*← *closestMap*.get(the identifier of entity *latestEnt*);
(o) If (*closestSet* == null) then *closestSet* ←new HashSet;
(p) *closestMap*.put(the identifier of entity *latestEnt*, *closestSet*);
(q) *closestSet* ← *closestSet*.add(the identifier of the i^{th} entity);
(r) } else if (*type_att*.equals("NUME") or *type_att*.equals("DATE")) Then {*winSize* ← narrow the window size according to (17);}
(s) *i* ←*i*++;
(t) If (*type_att*.equals("STRING") and *i* > *ws*_min) then recalculate the window size with (15);
(u) If (*i* > *winSize* and *similarity* > δ) then {*winSize* ←*winSize*++; Exit the current loop;//only for date or numerical}
(v) Sliding into the next window;
(w) return *closestMap*;

Algorithm 3: Calculate the similarity between entities in each block.

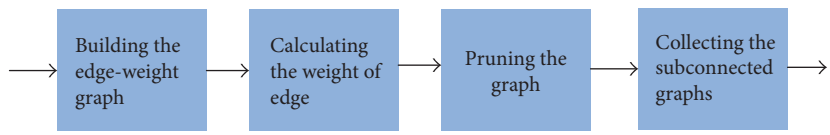


Figure 2: The internal functionality of our gathering method.

$$\text{ws\_}i = \begin{cases} \text{ws\_min} & (p \le \text{ws\_min}), \\ p & (\text{ws\_min} < p < \text{ws\_}i). \end{cases} \quad (17)$$

For the enumeration-type block, the window size is defined as follows:

$$\text{ws\_}i = \text{Max}(\text{frequency}(v\_i)), \quad (18)$$

where v_i represents the possible values of enumeration attribute.

3.3.3. MBN Algorithm Implementation. Based on the definitions of similarity metric in Section 3.3.1 and the variable window strategies in Section 3.3.2, we implement the closest entity detection in each block through Algorithm 3.

In order to obtain the final solution of the complete entities, we gather the output of each block obtained by Algorithm 3. Generally speaking, the gathering process is divided into the following four steps, which is intuitively represented as Figure 2.

(a) Building the edge-weight graph *G*_*e*, which is an undirected graph, is used to express *closestMap*. The nodes in this graph are the entity identifiers contained in the *closestMap*.

(b) If the likelihood of two entities matched is proved at least in one block, they will be connected by an edge. The purpose of this step is to determine the weight of each edge.

(c) Pruning edge-weight graphing aims at selecting the globally best pairs by iterating over the edges of an edge-weight graph in order to filter out edges that do not satisfy the pruning criterion, such as the edges with low weight.

(d) Collecting the connected subgraphs from the pruned edge-weight graph constitutes the final output of the complete entities.

In general, the strong indication of the similarity of two entities is provided by the number of blocks they have in common; the more blocks they share, the more likely they are to match. Therefore, the weight of an edge connecting entities r_i and r_j is set to the number of blocks which marked that these two entities are matched. However, the contribution of different blocks to the computation of the similarity between the complete entities is also different. For example, the output from a string-type block has a positive effect because their values are sufficient dispersion. As a result, we improved the common block scheme (CBS) method proposed in research [14, 35] and named it as improved common block scheme (ICBS), which is shown in (19). For the sake of discussion, we use e_i, j to describe the edge between entities r_i and r_j; correspondingly, e_i, j. weight is used to express its weight.

$$e\_i, j.\text{weight} = \sum_{k=1}^{d} e\_i, j^k.\text{weight} \times B\_i, j^k, \tag{19}$$

where e_i, j^k.weight is the weight of kth block, and if the entities r_i and r_j are marked as matched in the kth block, then $B\_i, j^k = 1$; else $B\_i, j^k = 0$. We adopt weight edge pruning (WEP) method [14] to prune the edges with lower weight.

4. Experiment Evaluation

This section aims to experimentally evaluate the techniques presented in this paper, respectively. We begin our analysis with the dataset collection, as shown in Section 4.1. In addition, the following experiments are conducted: the splitting attributes into different blocks and the attribute clustering methods are examined in Section 4.2. In order to evaluate the improved common block scheme (ICBS), we compare it with CBS presented in [14] in Section 4.3. At last, we compare our MBN method with the classical SNM in Section 4.4 to evaluate the effectiveness of our similarity metric schemes and dynamic adjustment window size strategies in different blocks. Among them, all approaches and experiments were fully implemented using Java 1.8 version, and development tool is IntelliJ IDEA. For the implementation of attribute clustering and pruning of edge-weight graph, we referred to and improved the source codes publicly released at http://SorceForge.net by Papadakis et al. [21]. We also employed some open source libraries in our implementation, such as jgrapht and commons-math.

4.1. Dataset Collection. To thoroughly evaluate our techniques, we employ a large-scale, real-world dataset, which is generated and obtained from one of real-world Internet of Things project that is developed and maintained by our research group. The time span of the dataset is from June 2016 to November 2016. Because if the terrorist can easily get flammable items (such as gasoline and natural gas), it will cause huge potential danger to the society. Therefore, the purpose of this project is to monitor these combustible materials in order to predict and monitor potential violent attacks in Xinjiang (one of China's provinces). In addition to refueling records, this project also collected information about drivers, vehicles, and so forth and stores them in the cloud platform. In combining machine learning algorithms, such as random forests or rotation forests, we will use these clean data output by our proposed method in this paper to propose some smart models in our future research, for example, to predict illegal vehicles.

4.2. Evaluation of Splitting Attributes to Different Blocks and Attribute Clustering Methods. For assessing the performance of our splitting attributes and attribute clustering methods, *Accuracy* is used as evaluation criteria in this part and its definition is shown as follows:

$$\text{Acc}_{\text{splitting}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\#\text{CorrectSplAttNum}}{\#\text{TotalAttNum}}, \tag{20}$$

$$\text{Acc}_{\text{clustering}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{\#\text{CorrectCluAttNum}}{\#\text{TotalAttNum}}, \tag{21}$$

where true positive (TP) is the amount of positive testing samples correctly predicted as positive, false positive (FP) is the amount of negative testing samples incorrectly predicted as positive, true negative (TN) is the amount of negative testing samples correctly predicted as negative, and false negative (FN) is the amount of positive testing samples wrongly predicted as negative.

In Table 1, in this experiment, we selected four tables to verify our methods. For the purpose of verifying the attribute clustering method, we added some redundant attributes and noise values in these four tables. The experimental results demonstrated that the average accuracy achieved by our splitting attributes to different blocks was 98.18% and the average accuracies obtained by our attribute clustering method from four tables were, respectively, 89.28%, 86.83%, 87.64%, and 92.08%. It can be proved that our methods work well. The main reason is that we adopt different methods to handle attributes based on the attribute value types and value distributions. For instance, in enumeration-type block, the confidence level was set as $\alpha = 0.001$, the similarity threshold was set to 0.80 in numerical- and date-type blocks, and it was set to 0.76 in string-type block.

4.3. Evaluation of ICBS Method. Combined with WEP [14] scheme, Figure 3 presents the accuracy comparison between CBS [14] and our ICBS. As shown in Figure 3, the experimental results exhibited that our ICBS had high efficiency in these four tables mentioned in Section 4.2 (i.e., accuracy > 84%). For ICBS, the weight in different type block

Table 1: The accuracy of splitting attributes to different blocking and attributes clustering.

Name	Attribute number	$Acc_{splitting}$	$Acc_{clustering}$			
			Enumeration	Date	String	Numerical
Cylinder_out	68	97.06%	86.67%	85.71%	92.6%	99.67%
Cylinder_check	69	98.8%	91.3%	89.47%	84.62%	85.71%
Vehicle_info	61	98.36%	87.5%	84.62%	90%	83.33%
Gas_cylinder	66	98.48%	91.67%	87.5%	83.33%	99.56%
Average	**66**	**98.18%**	**89.28%**	**86.83%**	**87.64%**	**92.08%**

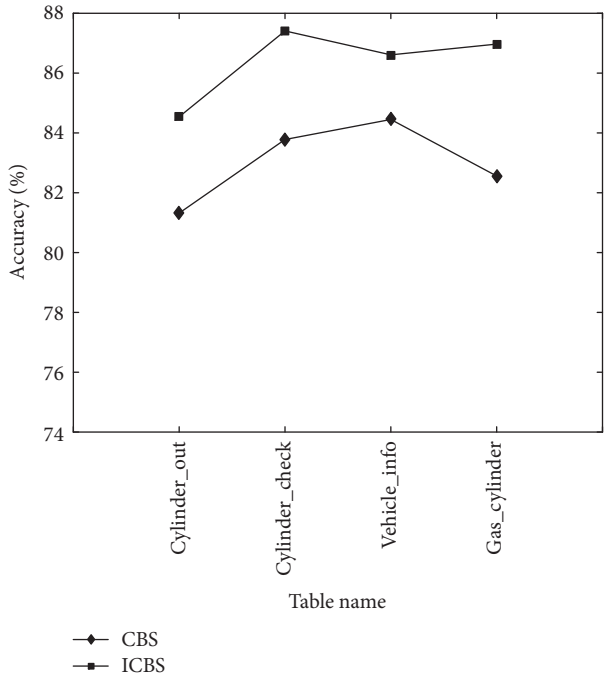


Figure 3: Accuracy comparison between ICBS and CBS.

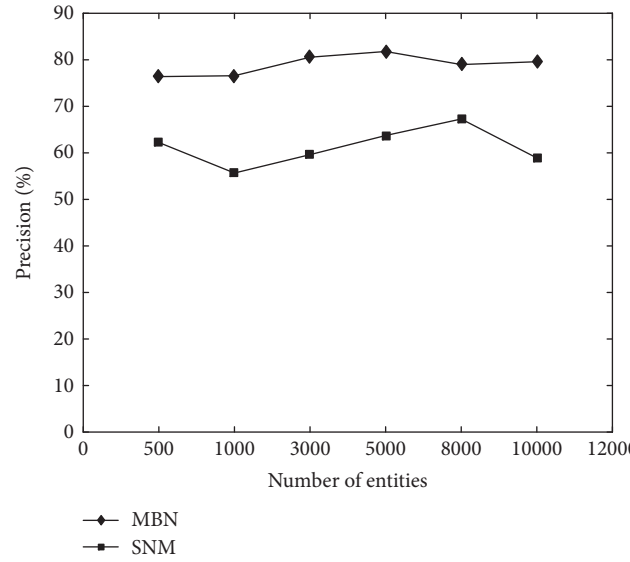


Figure 4: Precision comparison between MBN and SNM.

was also varied; generally speaking, the weight in string and numerical type was enlarged in these four tables, and the weight in enumeration and date was reduced. But for the CBS, the weight in different types of block was all set to 1. The experimental results demonstrated that these four types of block have different contributions to the calculation of the similarity of the complete entity and our ICBS is an appropriate and important part in our ER workflow.

4.4. Evaluation of MBN Method. In this section, we compared our MBN method with existing entity resolution technique SNM. The metrics are the following: (1) Precision (Pre.); (2) Recall (Rec.); (3) *F*-score, which are shown as follows:

$$\text{Pre.} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100\%, \tag{22}$$

$$\text{Rec.} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100\%, \tag{23}$$

$$F\text{-score} = \frac{2 \times \text{Pre.} \times \text{Rec.}}{\text{Pre.} + \text{Rec.}}, \tag{24}$$

where the TP and FP are the number of true-matched and true-nonmatched candidate entity pairs generated by the ER method. The FN is the number of nonmatched entity pairs.

We extracted the *Cylinder_check* table from the dataset to verify the effectiveness of our MBN method. This table has a total of 534,017 entities. We carried out this experiment based on five hundred, 1 thousand, 3 thousand, 5 thousand, 8 thousand, and 10 thousand entities, respectively. The entity similarity threshold was set to 0.79 both in MBN and SNM. The window size was set to 16 in SNM. In MBN, the initial window size for date- and numerical-type blocks was set to 8, the minimum window size was set to 5, the maximum window size was set to 30 for string type, and the window size of enumeration type was calculated in real-time. The input data of SNM and MBN were all from *Cylinder_check* table, in which the redundant attributes have been removed. The comparisons based on *precision*, *recall*, and *F-score* achieved by MBN and SNM are shown in Figures 4 and 5.

Figures 4 and 5 demonstrated that, compared with SNM, our MBN method improved the precision and recall for the closest entity detection. That means MBN is a more objective

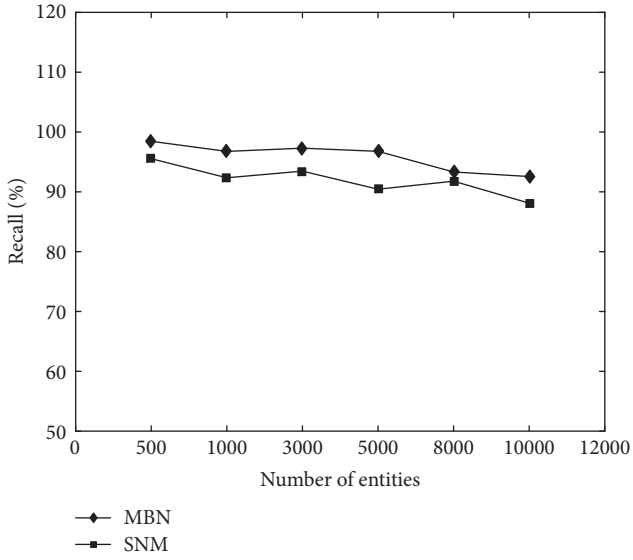


Figure 5: Recall comparison between MBN and SNM.

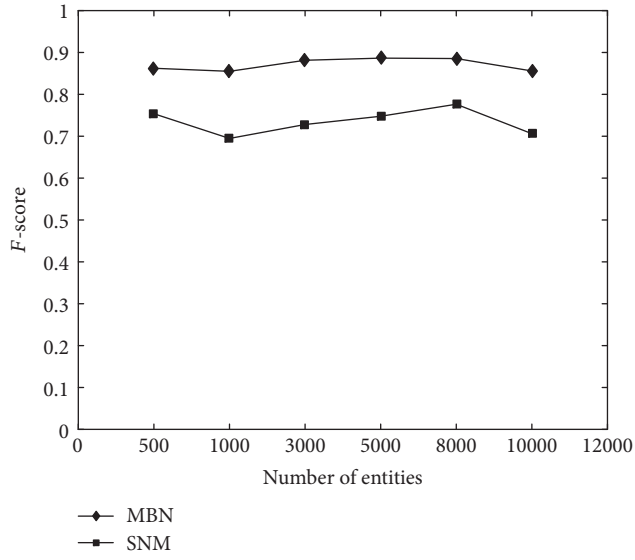


Figure 6: *F*-score comparison between MBN and SNM.

method to evaluate the similarity between entities because it split entities into different blocks based on the attribute value types and adopted different similarity metrics in these blocks. In addition, compared with the fixed window size in SNM, the dynamic adjustable window schemes introduced in our ER resolving method improves the performance while reducing the missing matched. Moreover, in SNM, the performance of detecting the closest entities mainly depends on its single sort key attribute, which means that if the key attribute are not selected properly or the key attribute contains noise data, its performance will be greatly reduced. Thus, it is necessary to select multiple key attributes and flexibly integrate several ER workflows to achieve better performance in MBN method. Generally speaking, these two methods all had acceptable recall under the same dataset. In Figure 6, the *F*-score comparison also indicated that MBN performed better than the original SNM method.

5. Conclusion and Future Work

In this paper, we introduce a novel hybrid approach, which is as an effective method for entity resolution in the context of voluminous and highly dimension data. In contrast to existing entity resolution methods, our approach commits to reducing the searching space for entity pairs by the constraint of redundant attributes and matching likelihood and to meeting multiple similarity comparison requirements based on attribute value types. Our thorough experimental evaluation verified the effectiveness, as well as the efficiency of our method on real-world dataset. We believe our approach is a promising alternative to entity resolution problem that can be effectively adopted to many other applications, such as the National Census area (e.g., it can be used to match data from different census collections to detect and correct conflicting or missing information) and Business Mailing Lists area (e.g., it can be used to identify redundant entities about the same customer to avoid money being wasted on mailing several copies of an advertisement flyer to one person). Of course, many interesting problems remain to be addressed in the future, including a general guidance for selecting key attribute in each block, and we will attempt to introduce parallelization technique into our approach to further improve its efficiency. Another interesting direction of our research is how to apply and develop our method to improve the efficiency of approaches that rely on entity resolution in machine learning and probabilistic inference.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported in part by the Xinjiang Key Laboratory Fund, under Grant no. 2016D03019, in part by the Xinjiang High-Tech R&D Program, under Grant no. 201512103, and in part by the Thousand Talents Plan, under Grant no. Y32H251201. The authors would like to thank all anonymous reviewers for their constructive advices.

References

[1] J. Wang, T. Kraska, M. J. Franklin, and J. Feng, "CrowdER: crowdsourcing entity resolution," *Proceedings of the Vldb Endowment*, vol. 5, no. 11, pp. 1483–1494, 2012.

[2] U. Draisbach and F. Naumann, "A generalization of blocking and windowing algorithms for duplicate detection," in *2011 International Conference on Data and Knowledge Engineering (ICDKE)*, Milan, Italy, 2011.

[3] G. Papadakis, E. Ioannou, and P. Fankhauser, "Efficient entity resolution for large heterogeneous information spaces," in *Proceedings of the fourth ACM international conference on Web search and data mining*, Hong Kong, China, 2011.

[4] J. Efremova, B. Ranjbar-Sahraei, H. Rahmani et al., "Multi-source entity resolution for genealogical data," in *Population*

Reconstruction, pp. 129–154, Springer International Publishing, Cham, 2015.

[5] S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina, "Entity resolution with iterative blocking," in *SIGMOD '09 Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, Providence, Rhode Island, USA, 2009.

[6] R. Mahmoud, N. El-Bendary, H. M. Mokhtar, and A. E. Hassanien, "Similarity measures based recommender system for rehabilitation of people with disabilities," in *Advances in intelligent systems and computing*, Springer, Cham, 2016.

[7] Y. R. Wang and S. E. Madnick, "The inter-database instance identification problem in integrating autonomous systems," in *[1989] Proceedings. Fifth International Conference on Data Engineering*, Los Angeles, CA, USA, 1989.

[8] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, pp. 39–48, Washington, D.C, 2003.

[9] P. Christen, "Automatic record linkage using seeded nearest neighbour and support vector machine classification," in *KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 151–159, Las Vegas, Nevada, USA, 2008.

[10] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Transactions on Knowledge & Data Engineering*, vol. 24, no. 9, pp. 1537–1555, 2012.

[11] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: a survey," *IEEE Educational Activities Department*, vol. 19, pp. 1–16, 2007.

[12] C. Xiao, W. Wang, X. Lin, J. X. Yu, and G. Wang, "Efficient similarity joins for near-duplicate detection," *ACM Transactions on Database Systems*, vol. 36, no. 3, p. 15, 2011.

[13] U. Draisbach and F. Naumann, "A comparison and generalization of blocking and windowing algorithms for duplicate detection," in *Proceedings of International Workshop on Quality in Databases (QDB)*, pp. 51–56, 2011.

[14] G. Papadakis, G. Koutrika, T. Palpanas, and W. Nejdl, "Meta-blocking: taking entity resolutionto the next level," *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 8, pp. 1946–1960, 2014.

[15] J. Z. Zhang, Z. Fang, Y. J. Xiong, and X. Y. Yuan, "Optimization algorithm for cleaning data based on SNM," *Journal of Central South University*, vol. 41, no. 6, pp. 2240–2245, 2010.

[16] L. Kolb, A. Thor, and E. Rahm, "Multi-pass sorted neighborhood blocking with MapReduce," *Computer Science - Research and Development*, vol. 27, no. 1, pp. 45–63, 2012.

[17] L. He, Z. Zhang, Y. Tan, and M. Liao, "An efficient data cleaning algorithm based on attributes selection," in *2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*, Seogwipo, South Korea, 2012.

[18] G. Chesi and K. Hashimoto, "Automatic image annotation approach based on optimization of classes scores," *Computing*, vol. 96, no. 5, pp. 381–402, 2014.

[19] Q. Tan, X. Zhang, and R. M. G. Kinshuk, "The 5R Adaptation Framework for Location-Based Mobile Learning Systems," *Modern Distance Education Research*, 2012.

[20] D. Srivastava, D. Srivastava, and D. Srivastava, "Less is more: selecting sources wisely for integration," in *International Conference on Very Large Data Bases*, 2012.

[21] G. A. Papadakis, *Blocking Techniques for Efficient Entity Resolution over Large, Highly Heterogeneous Information Spaces*, 2013.

[22] J. Sedano, S. González, C. Chira, Á. Herrero, E. Corchado, and J. R. Villar, "Key features for the characterization of android malware families," *Logic Journal of IGPL*, vol. 25, no. 1, pp. 54–66, 2016.

[23] J. R. Villar, P. Vergara, M. Menéndez, E. de la Cal, V. M. González, and J. Sedano, "Generalized models for the classification of abnormal movements in daily life and its applicability to epilepsy convulsion recognition," *International Journal of Neural Systems*, vol. 26, no. 06, article 1650037, 2016.

[24] J. Sedano, C. Chira, S. González, Á. Herrero, E. Corchado, and J. R. Villar, "Characterization of android malware families by a reduced set of static features," in *Advances in Intelligent Systems and Computing*, Springer International Publishing, Cham, 2016.

[25] Z. Q. Hao, Z. J. Zhang, and H. C. Chao, "A cluster-based fuzzy fusion algorithm for event detection in heterogeneous wireless sensor networks," *Journal of Sensors*, vol. 2015, Article ID 641235, 11 pages, 2015.

[26] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar, and S. Fienberg, "Adaptive name matching in information integration," *Intelligent Systems IEEE*, vol. 18, no. 5, pp. 16–23, 2003.

[27] A. Pomeroy and Q. Tan, "Effective SQL injection attack reconstruction using network recording," in *IEEE International Conference on Computer and Information Technology*, Pafos, Cyprus, 2011.

[28] S. Tejada, C. A. Knoblock, and S. Minton, "Learning domain-independent string transformation weights for high accuracy object identification," in *KDD '02 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, Edmonton, Alberta, Canada, 2002.

[29] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," *Kdd Workshop on Data Cleaning & Object Consolidation*, 2003.

[30] P. Christen, "A comparison of personal name matching: techniques and practical issues," in *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06)*, Hong Kong, China, 2006.

[31] A. Thor and E. Rahm, "Evaluation of entity resolution approaches on real-world match problems," *Proceedings of the VLDB Endowment*, vol. 3, pp. 484–493, 2010.

[32] Q. Tan, P. H. Wu, Y. L. Jeng, and Y. M. Huang, "Mobile computing architecture for multi-platform adaptation," in *Iadis Wireless Applications and Computing*, 2009.

[33] M. Ghantous and M. Bayoumi, "MIRF: a multimodal image registration and fusion module based on DT-CWT," *Journal of Signal Processing Systems*, vol. 71, no. 1, pp. 41–55, 2013.

[34] J. Han and M. Kamber, "Data mining: concepts and techniques," *Data Mining Concepts Models Methods & Algorithms Second Edition*, vol. 5, no. 4, pp. 1–18, 2006.

[35] Q. Tan and F. Pivot, "Big data privacy: changing perception of privacy," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Chengdu, China, 2015.

[36] A. Backurs and P. Indyk, "Edit distance cannot be computed in strongly subquadratic time (unless SETH is false)," in *STOC*

'15 Proceedings of the forty-seventh annual ACM symposium on Theory of computing, Portland, Oregon, USA, 2015.

[37] B. Karim, Q. Tan, I. El Emary, B. A. Alyoubi, and R. S. Costa, "A proposed novel enterprise cloud development application model," *Memetic Computing*, vol. 8, pp. 287–306, 2016.

[38] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *Proceedings 14th International Conference on Machine Learning (ICML'97)*, pp. 148–155, 1997.

[39] S. Loisel and Y. Takane, "Partitions of Pearson's chi-square statistic for frequency tables: a comprehensive account," *Computational Statistics*, vol. 31, no. 4, pp. 1429–1452, 2016.

[40] M. T. Puth, M. Neuhäuser, and G. D. Ruxton, "Effective use of Pearson's product–moment correlation coefficient," *Animal Behaviour*, vol. 93, pp. 183–189, 2014.

[41] D. Bogdanov, L. Kamm, S. Laur, P. Pruulmann-Vengerfeldt, R. Talviste, and J. Willemson, "Privacy-preserving statistical data analysis on federated databases," in *Lecture Notes in Computer Science*, pp. 30–55, Springer International Publishing, 2014.

[42] M. Denojean-Mairet, Q. Tan, F. Pivot, and M. Ally, "A ubiquitous computing platform - affordable telepresence robot design and applications," in *2014 IEEE 17th International Conference on Computational Science and Engineering*, Chengdu, China, 2014.

[43] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederee, and W. Nejdl, "A blocking framework for entity resolution in highly heterogeneous information spaces," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 12, pp. 2665–2682, 2013.

[44] M. A. Hernández and S. J. Stolfo, "The merge/purge problem for large databases," in *IGMOD '95 Proceedings of the 1995 ACM SIGMOD international conference on Management of data*, vol. 24, pp. 127–138, 1995.

[45] J. Ye, "Cosine similarity measures for intuitionistic fuzzy sets and their applications," *Mathematical and Computer Modelling*, vol. 53, no. 1-2, pp. 91–97, 2011.

[46] A. Abdelgawad and M. Bayoumi, "Proposed centralized data fusion algorithms," in *Lecture Notes in Electrical Engineering*, pp. 37–57, Springer US, Boston, MA, 2012.

[47] L. Muflikhah and B. Baharudin, "Document clustering using concept space and cosine similarity measurement," in *2009 International Conference on Computer Technology and Development*, pp. 58–62, Kota Kinabalu, Malaysia, 2009.

Hindawi
Journal of Sensors
Volume 2017, Article ID 4319709, 9 pages
https://doi.org/10.1155/2017/4319709

Research Article

Power Control for Passive QAM Multisensor Backscatter Communication Systems

Shengbo Hu,[1,2] Jinrong Mo,[1,2] Tingting Yan,[1,2] and Yanfeng Shi[1,2]

[1]*Institute of Intelligent Information Processing, Guizhou Normal University, Guiyang 550001, China*
[2]*Center for RFID and WSN Engineering, Department of Education, Guizhou Normal University, Guiyang 550001, China*

Correspondence should be addressed to Jinrong Mo; 935472997@qq.com

Received 25 April 2017; Accepted 3 October 2017; Published 27 December 2017

Academic Editor: Nashwa El-Bendary

To achieve good quality of service level such as throughput, power control is of great importance to passive quadrature amplitude modulation (QAM) multisensor backscatter communication systems. First, we established the RF energy harvesting model and gave the energy condition. In order to minimize the interference of subcarriers and increase the spectral efficiency, then, the colocated passive QAM backscatter communication signal model is presented and the nonlinear optimization problems of power control are solved for passive QAM backscatter communication systems. Solutions include maximum and minimum access interval, the maximum and minimum duty cycle, and the minimal RF-harvested energy under the energy condition for node operating. Using the solutions above, the maximum throughput of passive QAM backscatter communication systems is analyzed and numerical calculation is made finally. Numerical calculation shows that the maximal throughput decreases with the consumed power and the number of sensors, and the maximum throughput is decreased quickly with the increase of the number of sensors. Especially, for a given consumed power of sensor, it can be seen that the throughput decreases with the duty cycle and the number of sensors has little effect on the throughput.

1. Introduction

Since the advent of backscatter communication, it has been widely used in the passive RFID (radio frequency identification) system due to its low cost and low power [1, 2]. Similarly, using the principle of the backscatter communication, the passive sensor node transmitter can be simplified into a transistor connected to the antenna. This will greatly reduce the cost and power consumption of wireless communication systems. For example, scholars from the University of Washington [3] researched the WISP (wireless identification and sensing platform) wireless smart sensor systems which can harvest RF energy based on backscatter communication. Now, PBC (passive backscatter communication) has been applied in these fields of smart car, wireless temperature measurement, the biological signal acquisition [4–9], and so forth. Undoubtedly, the study and application of passive backscatter communication will attract more and more attention.

In passive backscatter communication systems, sensor nodes need to harvest RF energy from the transceivers or reader and use it to recharge their finite energy storage capacity as shown in Figure 1. However, the unpredictable environments such as the channel, the number of sensors, and challenges of the RF energy harvesting make the activities of the sensors difficult, including sensing, processing, and communicating by nodes with a given rate. So, an efficient energy management is necessary in order to guarantee sensors' activities and a QoS (quality of service) of backscatter communication [10–12].

In fact, the mode using RF energy harvesting as a supplement to the finite energy storage capacity is called energy neutral operation [13]. In this mode, for such RF energy harvesting nodes with ideal finite energy storage capacity, the condition for energy neutral operation should be satisfied for all nonnegative values of τ:

$$\int_0^\tau P_c(t)dt \le \int_0^\tau P_{RF}(t)dt + B_0, \quad (1)$$

where $P_{RF}(t)$ is the harvested RF power at time t, $P_c(t)$ is the consumed power by nodes at that time, B_0 is the initial

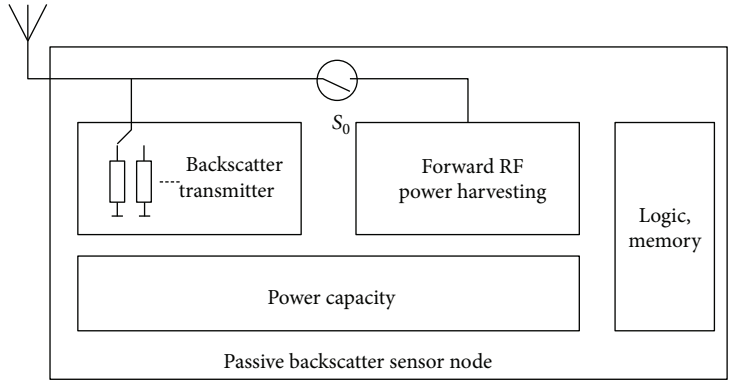


Figure 1: Passive backscatter sensor node.

energy stored in the ideal capacity, and τ is the runtime. So, the goal of power control for passive backscatter communication is to maximize QoS level under condition of energy neutral operation.

For implementing multisensor backscatter communication systems, subcarrier modulation can be used for each sensor [14]. This leads to interference among sensors. In order to minimize such interference and increase spectral efficiency, modulations including QAM are used to improve data throughput [15–17]. It shows that it is feasible to use load-dependent scattering for yielding a simple QAM backscatter [15, 17]. In this paper, we focus on colocated multisensor QAM backscatter communication, showing that power control is important for the backscatter communication systems to achieve good QoS level via analysis and simulation. We introduce a related work in Section 2 and describe a system model in Section 3. In Section 4, we present the performance of the backscatter communication systems. And we analyze the impact of power control on the throughput of systems in Section 5. Finally, we provide some concluding remarks in Section 6.

2. Related Work

Several papers have mainly proposed power control on nodes using energy harvesting, because power control of energy harvesting determines the level of connectivity of backscatter communication as well as the achievable QoS of the systems. An RF-powered transmitter that supports 915 MHz downlink and 2.45 GHz uplink bands is designed in [18]. In [19], the paper computes the minimal number of sinks required to keep the network connected and to satisfy the required constraints.

There exist attractive applications of the RF-powered devices such as wireless body networks. Benefiting from RF energy harvesting, some low-power devices can achieve real-time work-on-demand power from RF sources, which further enables a battery-free circuit with reduced size [20]. The body devices that implement high efficiency can be found in [21, 22]. For the multisensor system, we can refer [23]. The authors address the problem of developing energy-efficient transmission strategies for body sensor networks with energy harvesting [24]. And the authors evaluate the impact of transmit power control on the usefulness of a multisink WSN-Heap using energy harvesting, deployed in uniform string topology for railway track monitoring [25].

3. System Models

3.1. RF Energy-Harvesting Model. The RF energy harvester is composed of an antenna, impedance matching circuit, rectifier, and voltage multiplier as shown in Figure 2. The efficiency η of the RF energy harvester is defined as follows:

$$\eta = \eta_1 \eta_2 \eta_3 = \frac{P_{\text{eh}}}{P_{\text{in}}}, \tag{2}$$

where η_1, η_2, and η_3 are the efficiencies of the receiving antenna, impedance transformation network, and the rectifier, respectively, P_{eh} is the output DC power and the meaning of eh is energy harvesting, and P_{in} is the input RF power.

In Figure 2, Y-stage voltage multiplier circuits are used to promote the output voltage. So, the upper limit of the steady-state output voltage V_{dd} is increased by a factor Y and can be written as [26]

$$V_{\text{dd}} = 2Y\left(V_{\text{peak}} - V_{\text{on}}\right), \tag{3}$$

where V_{peak} is the peak voltage of the signal at the input of the view of the Y-stage voltage multiplier circuits and V_{on} is the turn-on voltage of the diode.

Now, let us consider the minimum voltage for nodes to operate. Assuming, the harvesting power from the transceivers or reader is $P_{\text{eh}}(t)$ and the consumed power by nodes at that time is $P_{\text{c}}(t)$. When the power made available by the transceivers or reader is insufficient for continuous operation, that is, $P_{\text{eh}}(t) < P_{\text{c}}(t)$, a capacity must be used to store RF energy until enough RF energy exists to complete an access. This affects the backscatter communication QoS level. Under these circumstances, the capacitor must be charged unless the stored energy E_{stored} is greater than or equal to the access energy E_{access}, where E_{access} is the energy required for each access.

The stored energy E_{stored} can be written as

$$E_{\text{stored}} = \frac{1}{2} C\left(V_{\text{dd}}^2 - V_{\text{node}}^2\right), \tag{4}$$

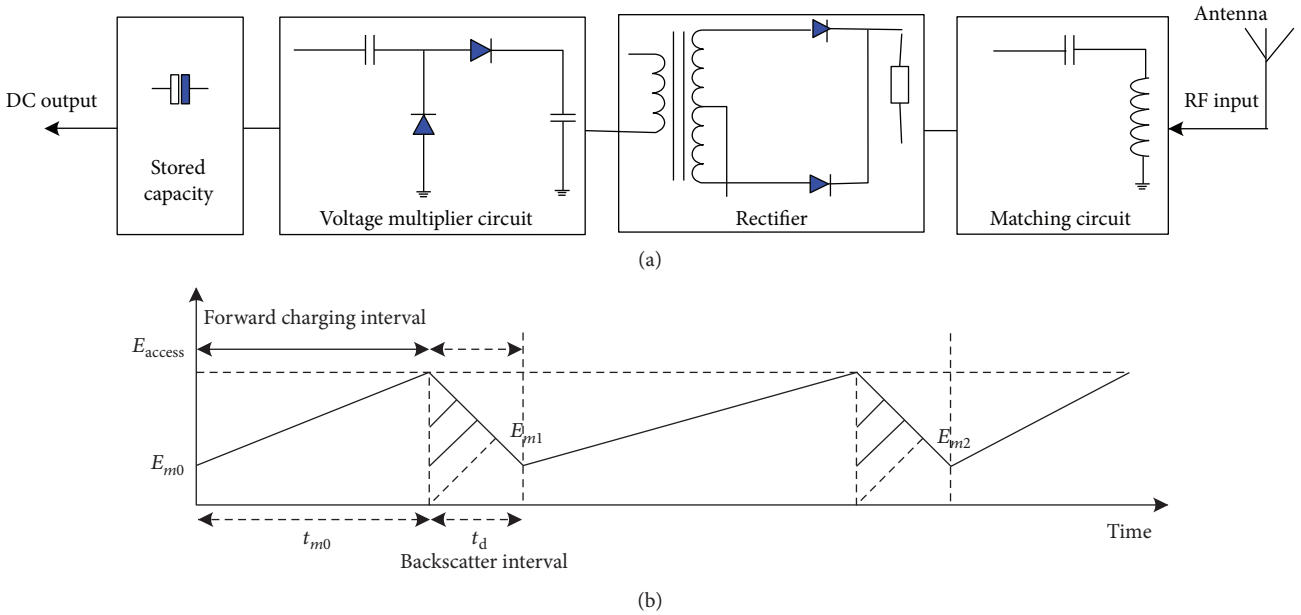


Figure 2: RF energy-harvesting model.

where C is the value of energy harvesting circuit output capacity and V_{node} is the voltage level at which the node operates. And the access energy E_{access} can be written as

$$E_{access} = \int_{T_0}^{T_0+T_{access}} P_c dt = V_{node} \int_{T_0}^{T_0+T_{access}} I_{access} dt, \quad (5)$$

where $[T_0, T_0 + T_{access}]$ is the access interval and I_{access} is the time-varying current draw of the node during each access.

Hence, the minimum voltage for nodes to operate is solved as

$$V_{dd}^{min} = \sqrt{\left(\frac{2E_{access}}{C}\right) + V_{node}^2}. \quad (6)$$

Once $V_{dd} < V_{dd}^{min}$, the node can not operate.

3.2. Energy Condition for Node Operating. Backscatter communication links include forward link and backscatter link. While accessing, the node harvests RF energy from the transceivers or reader firstly by forward link. Once $V_{dd} \geq V_{dd}^{min}$, that is, $E_{stored} \geq E_{access}$, the node draws power to carry out its operation and sends the sensor data to the transceivers or reader by backscatter link. These can be defined as an energy model as given in Figure 2. Each access interval comprises a charging phase and backscatter communication phase. In Figure 2, while accessing at time T_0, the capacitor C charges unless the stored energy is $E_{stored} \geq E_{access}$ at time $T_0 + T_c$. Then, the node sends data by backscatter link. Once $E_{stored} < E_{access}$ at time $T_0 + T_c + T_d$, the node stops operating and waits next access. Let $E_{eh} = B_0 + \eta \int_{T_0}^{T_0+T_c} [P_{eh}(t) - P_c(t)]^+ dt$, the condition for energy neutral operation should be satisfied in the access interval $[T_0, T_0 + T_{access}]$:

$$E_{eh} - \int_{T_0+T_c}^{T_0+T_c+T_d} [P_c(t) - P_{eh}(t)]^+ dt \geq 0, \quad (7)$$

where $\eta \in [0, 1]$ is the charging rate, B_0 is the initial energy stored in the capacity, and the rectifier function $[x]^+$ is defined as follows:

$$[x]^+ = \begin{cases} x, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (8)$$

Meanwhile, the limitation of capacity size requires the constraint to be satisfied as follows:

$$E_{eh} - \int_{T_0+T_c}^{T_0+T_c+T_d} [P_c(t) - P_{eh}(t)]^+ dt \leq B, \quad (9)$$

where B is the maximum energy that can be stored in the capacitor.

Hence, the energy conditions for the node operating include (7), (9), and (10). And (10) is given as follows:

$$B_0 + \eta \int_{T_0}^{T_0+T_c} [P_{eh}(t) - P_c(t)]^+ dt \geq E_{task}, \quad (10)$$

where E_{task} is the minimum energy required for the node to operate.

For simplicity, assuming the consumed power by the node with forward link is P_{fc}, the consumed power by the sensor with backscatter link is $P_{bc}(P_{fc} < P_{bc})$ and P is

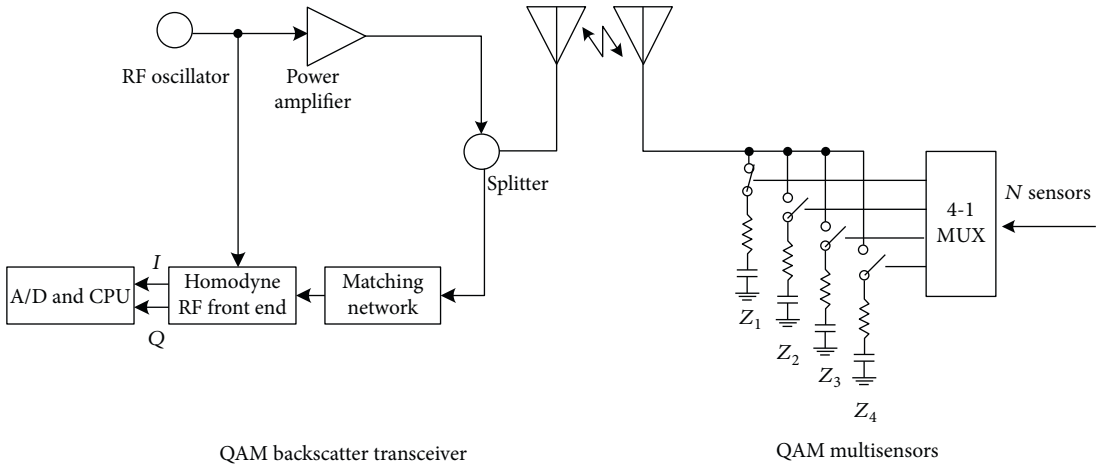


Figure 3: Passive QAM backscatter multisensor communication systems.

the harvested power, then the energy condition for node operating is given as follows:

$$\begin{aligned} B_0 + \eta T_c(P - P_{fc}) - T_d P_{bc} &\geq 0, \\ B_0 + \eta T_c(P - P_{fc}) - T_d P_{bc} &\leq B, \\ B_0 + \eta T_c(P - P_{fc}) &\geq E_{access}. \end{aligned} \tag{11}$$

4. Performance Analysis

4.1. Colocated QAM Multisensor Backscatter Signal. Typically, backscatter communication systems use single-frequency continuous wave transmission waveforms for accessing. The resulting complex signal from the node received at the transceivers or reader is given as

$$r(t) = \sqrt{2E\alpha} m(t) \exp[j(\omega_c t + \varphi)] + n(t), \tag{12}$$

where E, ω_c, and φ are the peak power, the angular frequency, and the phase of the carrier signal, respectively, $n(t)$ is AWGN (additive white Gaussian noise), and $m(t)$ is the message signal to be transmitted.

To improve spectral efficiency, the use of QAM opens up many new avenues for backscatter communication systems. It has been shown recently that it is feasible to build backscatter systems supporting QAM using node load selection methodology [17, 27]. However, improved spectral efficiency means that the node's impedance is mismatched and the power harvested or the efficiency of impedance transformation network η_2 is reduced.

What is more, $m(t)$ can be given as

$$m(t) = \sum_{k=-\infty}^{\infty} a_k g(t - kT), \tag{13}$$

where $g(t)$ is the rectangle pulse with width T and $\{a_k\}$ is the sequence of symbols with L levels. If the number of phases φ is K in (12), the number of the carrier states is $K \times L$. And when $K = L = 2$, the modulation is 4QAM, that is, to design and implement 4QAM backscatter modulator, four RC lumped impedances are connected to an antenna port through RF 4-1 Mux, each lumped impedance corresponding to different reflection coefficients. The design of 4QAM colocated backscatter communication systems with this procedure is given in Figure 3. In Figure 3, serial/parallel converts sensor data into 2 bits for controlling the 4-1 Mux.

To make the structure of the backscatter multisensor nodes simple, we put forward an access scheme with different subcarrier modulations for each sensor; namely, each sensor has different pulse width. So, the cumulative colocated QAM multisensor backscatter signal at the transceivers or reader is a sum of all complex signals from sensors with additive white noise, resulting in

$$\begin{aligned} r(t) = {} & \sqrt{2P_i\alpha_i} m_i(t) \exp[j(\omega_c t + \varphi)] \\ & + \sum_{k \neq i} \sqrt{2P_k\alpha_k} m_k(t) \exp[j(\omega_c t + \varphi)] + n(t). \end{aligned} \tag{14}$$

4.2. Throughput Maximization for Colocated QAM Multisensor under the RF Energy-Harvesting Constraints. The goal of power control for passive multisensor random interrogating is to meet some QoS criteria, for example, throughput.

In Figure 3, each sensor has different subcarrier frequencies; however, the information leak of one sensor interferes other sensors. The amount of such interference depends on the power spectrum of the chosen subcarrier modulation and the number of sensors. According to the constraint of the correlated short-range wireless communication spectrum, the power spectrum of QAM modulation of the ith sensor can be approximated as [28]

$$S(f) \approx \frac{1}{(f - f_i)^4}, \tag{15}$$

where $f_i = 1/T_i$ and T_i is the pulse width of the ith sensor data.

Under the lognormal channel, signal-to-interference-and-noise ratio (SINR) at the transceivers or reader in Figure 3 is written as

$$\Gamma_i = \frac{2E_i h_i}{(W_0/T_i) + \sum_{j=1,\neq i}^{N} 2E_j h_j S\left(f_j - f_i\right)}, \tag{16}$$

where E_i is the transmitted bit energy, W_0 is the power spectrum density, $h_i = l_i \Omega_i$, l_i is the path loss, Ω_i is the shadow fading of logarithmic normal distribution, and $\Omega_i = e^x$. x follows normal distribution $G(\mu, \sigma^2)$.

In this paper, the multiple sensors are colocation and share a transmitter and transceiver, so it can be assumed that $h_i = h_j$ and $E_i = E_j$. Consequently, (16) can be rewritten as

$$\Gamma_i = \frac{1}{\Gamma_N^{-1} + \Gamma_I^{-1}}, \tag{17}$$

where $\Gamma_N = 2E_i T_i \cdot l_i^{\gamma_i} \cdot 10^{x_i/10}/W_0$ and $\Gamma_I = \sum_{j=1,\neq i}^{N} 10^{(x_i - x_j)/10} \cdot (l_i/l_j)^{\gamma} \cdot (1/(S(f_j - f_i)))$.

Now, we characterize the performance of backscatter communication. We start by defining the concept of the throughput.

Definition 1. The data delivery ratio D_r of backscatter communication systems is a ratio of successful data packet received to attempted data packet transmitted.

Definition 2. The throughput T_h of backscatter communication systems is the probability that a packet is successfully received during a data interrogation interval.

For a packet to be received successfully, the SINR at the transceivers or reader must exceed some threshold λ_{th}. Using the definitions above, the data delivery ratio D_{ri} and the throughput T_{hi} in the access interval T_{access} of the ith sensor can be written, respectively, as

$$\begin{aligned} D_{ri} &= P\{\Gamma_i \geq \lambda_{th}\}, \\ T_{hi} &= \frac{P\{\Gamma_i \geq \lambda_{th}\}}{T_{access}}. \end{aligned} \tag{18}$$

So, the average data delivery ratio and throughput of the backscatter communication systems can be written as follows:

$$\begin{aligned} D_r &= \frac{\sum_{i=1}^{N} D_{ri}}{N}, \\ T_h &= \frac{\sum_{i=1}^{N} T_{hi}}{N}, \end{aligned} \tag{19}$$

where N is the number of sensors.

Generally, obtaining throughput in a closed form is not always possible and remains analytically challenging; because of the cumulative distribution function of random variable, Γ_i remains difficult. Thanks to nodes colocating, this problem can be solved easily. In this case, $x_i = x_j$, $l_i = l_j$, and (16) can be rewritten as

$$\Gamma_i = \frac{1}{\Gamma_N^{-1} + \sum_{j=1,\neq i}^{N} S\left(f_j - f_i\right)}. \tag{20}$$

Let

$$k_i = 10\log\left(\frac{W_0}{2E_i T_i l_i^{\gamma_i}} \cdot \frac{\lambda_{th}}{1 - \lambda_{th} \sum_{j=1,\neq i}^{N} S\left(f_j - f_i\right)}\right), \tag{21}$$

then,

$$\begin{aligned} D_{ri} &= P\{\Gamma_i \geq \lambda_{th}\} = P\{x_i \geq k_i\} \\ &= 1 - P(x_i < k_i) = 1 - \int_{-\infty}^{k_i} \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\left(x_i^2/2\sigma_i^2\right)} dx. \end{aligned} \tag{22}$$

Hence, the average data throughput T_h can be rewritten as

$$T_h = \frac{\sum_{i=1}^{N} D_{ri}}{N T_{access}} = \frac{\sum_{i=1}^{N} \left(1 - \int_{-\infty}^{k_i} \left(1/\sqrt{2\pi}\sigma_i\right) e^{-\left(x_i^2/2\sigma_i^2\right)} dx\right)}{N T_{access}}. \tag{23}$$

Assuming duty cycle $\rho = \bar{T}_d/(\bar{T}_c + \bar{T}_d) = \bar{T}_d/\bar{T}_{task}$, $\rho \in [0, 1]$, the nonlinear optimization problems of power control for passive QAM backscatter communication systems with N nodes can be expressed as

$$\max T_h = \frac{\sum_{i=1}^{N} D_{ri}}{N T_{access}}, \tag{24}$$

subject to

$$\begin{aligned} B_{0m} + \eta(1-\rho)\bar{T}_{task}(P - NP_{fc}) - \rho N \bar{T}_{task} P_{bc} &\geq 0, \\ B_{0m} + \eta(1-\rho)\bar{T}_{task}(P - NP_{fc}) - \rho N \bar{T}_{task} P_{bc} &\leq B, \\ B_{0m} + \eta(1-\rho)\bar{T}_{task}(P - NP_{fc}) &\geq E_{task}, \\ B_0 &\leq B. \end{aligned} \tag{25}$$

5. Numerical Results and Discussions

We evaluate the performance of the power control using numerical analysis. The analysis includes the impact of P_{bc} on the throughput and the harvested power.

5.1. Parameters for Numerical Analysis. Referring to the relation between bit error rate and SINR for 4QAM under AWGN channel, the threshold λ_{th} can be chosen as 3.0156 corresponding to the data delivery ratio ranging from 1 to 1×10^{-5}. Meanwhile, the given transmission data rate is 40 kbps, the data packet size is 64 bits, and the duration of a data packet transmission is $T_{dm} = 1.6$ sec. All parameters used are listed in Table 1 [29].

5.2. Impact of P_{bc} on Throughput T_h. In this section, we first investigate the choice of P_{bc} to achieve maximal throughput.

Table 1: Parameters for numerical analysis.

Parameter	Value
W_0	−1.0779 dBw/Hz
l_i	−1 dB
λ_{th}	3.0156
T_{dm}	1.6 sec
P_{fc}	68.9 nW
B	0.0125 J
η	0.65

Under the energy condition for node operating in (11) and solving the inequality group, it can be shown that the maximum and minimum access interval T_{access} is given by

$$T_{\text{access}}^{\min} = \frac{\eta T_d (P - NP_{\text{fc}}) + NT_d P_{\text{bc}}}{\eta (P - NP_{\text{fc}})},$$
$$T_{\text{access}}^{\max} = \frac{\eta T_d (P - NP_{\text{fc}}) + NT_d P_{\text{bc}} + B}{\eta (P - NP_{\text{fc}})}. \quad (26)$$

The access interval T_{access} is a function of P_{bc}. When $P = 189\,\mu W$, $N = 2$, $N = 4$, and $N = 8$, the minimum and maximum access intervals of sensor node1 with subcarrier $f_1 = 50000$ Hz are plotted in Figures 4 and 5. The subcarriers of other nodes are 50,002 Hz, 50,010 Hz, 50,100 Hz, 50,110 Hz, 50,120 Hz, 50,150 Hz, and 50,200 Hz (the same below). According to these frequencies and the value of T_{access}, we can deduce the corresponding value of the throughput using (20), (23), and (24). And the throughput changes by varying of T_{access}.

It is observed that the maximum and minimum access intervals T_{access} increase with P_{bc} in Figures 4 and 5. It is easy to understand, because the greater the P_{bc}, the longer the charge time to meet the energy condition for node operating. It is worthwhile to note that the maximum and minimum access interval T_{access} increases more with the increase of N.

Hence, the maximum throughput as a function of P_{bc} is plotted under the condition of the minimum access interval T_{access} in Figure 6. It is found that the maximal throughput decreases with the increase of P_{bc}. And the maximum throughput is decreased quickly with the increase of N.

5.3. Impact of Duty Cycle ρ on Throughput T_{h}. Next, we investigate the impact of varying duty cycle ρ on the throughput T_{h}.

Similar to Section 4.2, the maximum and minimum duty cycle ρ under the energy condition for node operating is given as follows:

$$\rho_{\max} = \frac{\eta T_d (P - NP_{\text{fc}})}{\eta T_d (P - NP_{\text{fc}}) + NT_d P_{\text{bc}}},$$
$$\rho_{\min} = \frac{\eta T_d (P - NP_{\text{fc}})}{B + \eta T_d (P - NP_{\text{fc}}) + NT_d P_{\text{bc}}}. \quad (27)$$

For a given $P_{\text{bc}} = 0.524\,\mu$W and a given $P_{\text{fc}} = 0.0689\,\mu$W, the throughput T_{h} as a function of duty cycle ρ is shown in Figure 7. It can be seen that the throughput T_{h} decreases with the increase of duty cycle ρ. Because decreasing duty cycle ρ leads to the increasing of the access interval T_{access}.

It should be noted that N has little effect on the throughput in Figure 7, because duty cycle ρ is insensitive to N. Especially, N has almost no effect on the minimum duty cycle ρ. We can analyze this using parameter sensitivity.

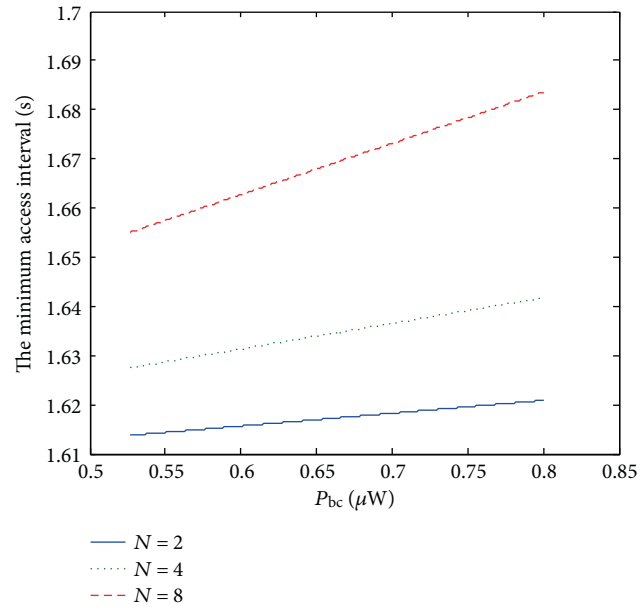


Figure 4: The minimum access interval versus the consumed power by sensor with scatter link.

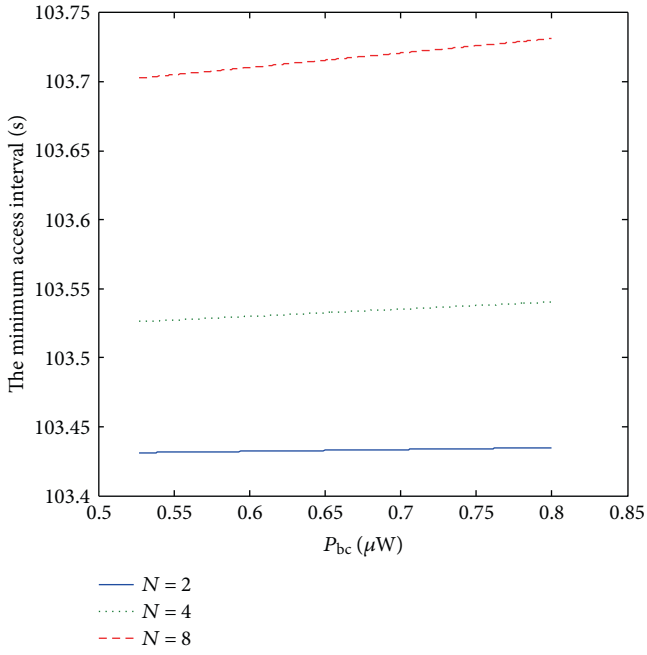


Figure 5: The maximum access interval versus the consumed power by sensor with scatter link.

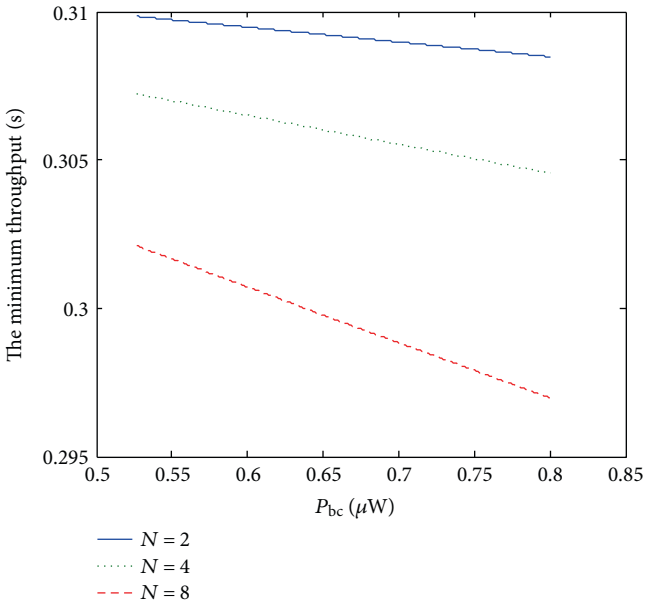


Figure 6: The maximum throughput versus the consumed power by sensor with scatter link.

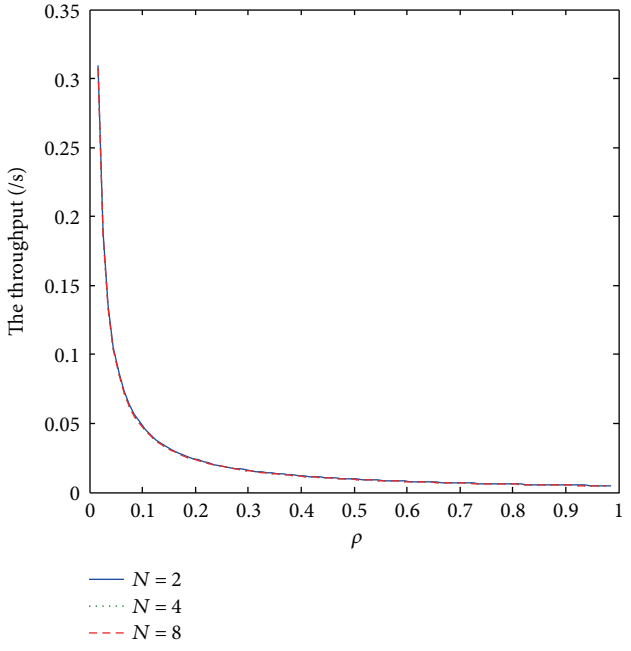


Figure 7: The throughput versus the duty cycle ρ.

Taking the partial derivative, the sensitivity of $\rho_{\max}$ and $\rho_{\min}$ is given as

$$S_{\rho_{\max}} = \left|\frac{\partial \rho_{\max}}{\partial N}\right| = \frac{\eta T_d^2 P P_{bc}}{\left[\eta T_d (P - N P_{fc}) + N T_d P_{bc}\right]^2},$$
$$S_{\rho_{\min}} = \left|\frac{\partial \rho_{\min}}{\partial N}\right| = \frac{\eta \left(T_d^2 P P_{bc} + T_d B P_{fc}\right)}{\left[B + \eta T_d (P - N P_{fc}) + N T_d P_{bc}\right]^2}. \quad (28)$$

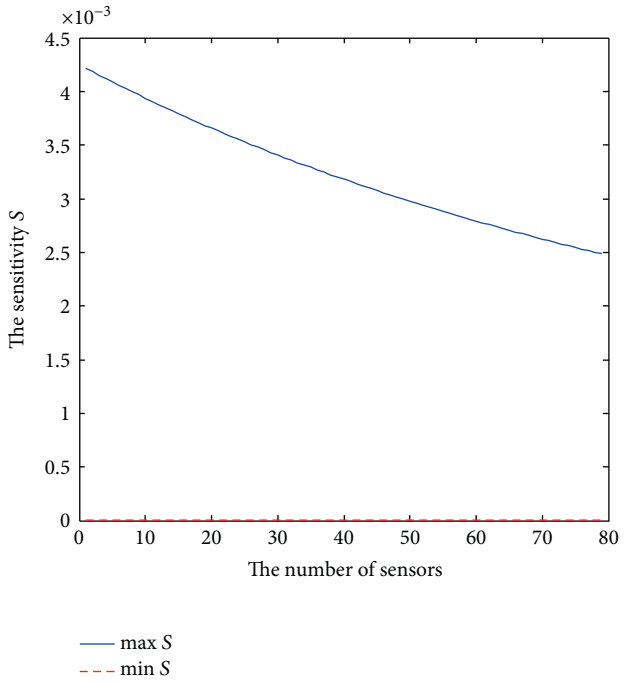


Figure 8: The sensitivity of duty cycle about N.

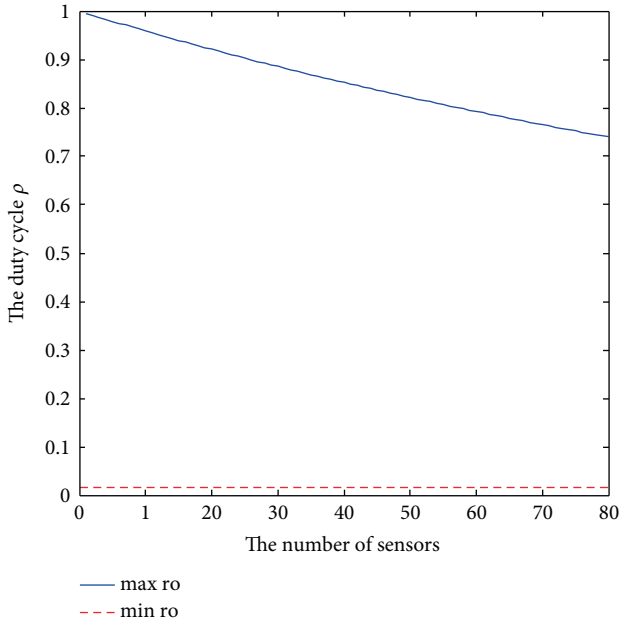


Figure 9: The relation between the duty cycle ρ and N.

Hence, the $S_{\rho_{\max}}$ and $S_{\rho_{\min}}$ as functions of N are shown in Figure 8. In Figure 8, $S_{\rho_{\max}}$ and $S_{\rho_{\min}}$ are very small. Especially, N has almost no effect on the minimum duty cycle ρ, because of using the large stored capacity. This can be also seen from the relation between duty cycle ρ and N shown in Figure 9.

5.4. Impact of P_{bc} on the Minimum Harvested Energy P_{min}. Finally, we analyze the choice of P_{bc} to achieve

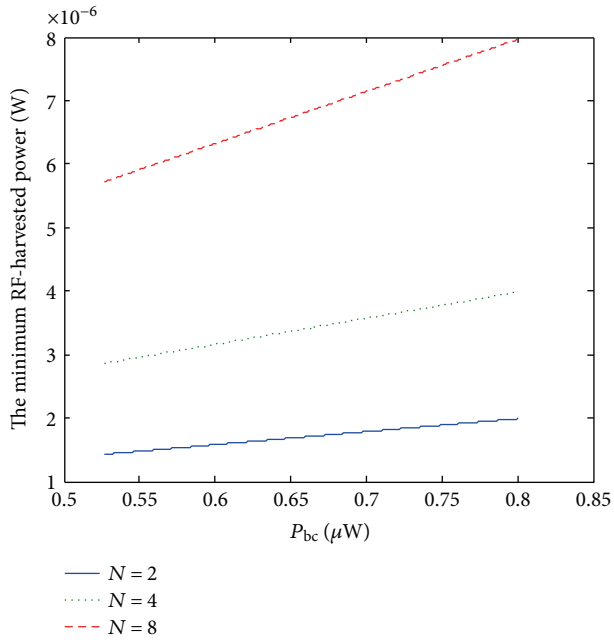


FIGURE 10: The minimum RF-harvested power versus the consumed power by sensor with scatter link.

the minimal RF-harvested energy under the energy condition for node operating.

In the same way, we can derive that the minimal RF-harvested energy is given by

$$P_{\min} = NP_{bc} + \frac{E_{task}}{\eta(1-\rho)T_{task}}. \quad (29)$$

The $P_{\min}$ as a function of P_{bc} is plotted in Figure 10. In Figure 10, as the P_{bc} and N increase, the $P_{\min}$ increases. This conclusion is obvious. Similar to Section 5.2, the number of sensors has larger effect on the minimum RF-harvested power.

6. Conclusion

To achieve good QoS level such as throughput, power control is of great importance to passive QAM multisensor backscatter communication systems. This paper presents the RF energy-harvesting model and gives the energy condition for node operating. In order to minimize interference resulted from multisensors and increase spectral efficiency, we propose the colocated passive QAM backscatter communication signal model, give the closed-form solution of the throughput under log-normal channel, and put forward the nonlinear optimization problems of power control for passive QAM backscatter communication systems. Solving the nonlinear optimization problems, we obtain the maximum and minimum access intervals, the maximum and minimum duty cycle, and the minimal RF harvested energy under the condition of node operating. Based on the solutions above, we analyze the maximum throughput of passive QAM backscatter communication systems and make numerical calculation. Finally, we make the conclusion as follows:

(i) The energy condition for node operating is a prerequisite for designing and optimizing the passive backscatter communication systems, and the goal of power control is to maximize the QoS of the systems.

(ii) Under the energy condition for node operating, the consumed power P_{bc} of the sensor with backscatter link and the number of sensors has large effect on the throughput of systems. It is found that the maximal throughput decreases with the increase of P_{bc} and N and the maximum throughput is decreased quickly with the increase of N.

(iii) Especially, for a given P_{bc} and a given P_{fc}, it can be seen that the throughput T_h decreases with the increase of duty cycle ρ and the number of sensors has little effect on the throughput.

(iv) There also exists a further improvement in this paper. For the next phase work, we will try to establish an actual model for data measurement and we will consider the WCDMA or LTE signals substituting for QAM signals.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research is a project partially supported by the National Natural Science Foundation of China (Grant no. 61362004) and Guizhou Joint Natural Science Foundation (Grant no. LKS [2013] 25).

References

[1] H. Stockman, "Communication by means of reflected power," *Proceedings of the IRE*, vol. 36, no. 10, pp. 1196–1204, 1948.

[2] A. Sample, D. Yeager, and P. Powledge, "Design of an RFID-based battery-free programmable sensing platform," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 11, pp. 2608–2615, 2008.

[3] D. Yeager, F. Zhang, and A. Zarrasvand, "A 9μA, addressable Gen2 sensor tag for biosignal acquisition," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 10, pp. 2198–2209, 2010.

[4] W. U. Bajwa, A. Sayeed, and R. Nowak, "Matched source-channel communication for field estimation in wireless sensor networks," in *Proceedings of Information Processing in Sensor Networks, 2005*, Los Angeles, CA, USA, April 2005.

[5] S. Ergen, A. S. Vincentelli, and X. Sun, "The tire as an intelligent sensor," *IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems*, vol. 28, no. 7, pp. 941–955, 2009.

[6] J. Grosinger, "Feasibility of backscatter RFID systems on the human body," *EURASIP Journal on Embedded Systems*, vol. 2013, no. 1, pp. 1–10, 2013.

[7] S. Hu, S. Bing, and S. Heng, "Power transmission performance for ultra high frequency embedded RFID system in tire," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 29, no. 21, pp. 150–158, 2013.

[8] P. V. Bharat, N. Sihna, and K. E. Pujitha, "Tire pressure monitoring system using ambient backscatter technology containing RF harvesting circuitry," *International Journal of Advance Engineering and Research Development*, vol. 1, no. 6, pp. 1–11, 2014.

[9] P. Ferrari, A. Flammini, and E. Sisinni, "New architecture for a wireless smart sensor based on a software defined radio," *IEEE Transaction on Instrumentation and Measurement*, vol. 60, no. 6, pp. 2133–2141, 2011.

[10] A. Kansal, D. Potter, and M. B. Srivastava, "Performance aware tasking for environmentally powered sensor networks," in *Proceedings of ACM SIGMETRICS*, vol. 32, no. 1, pp. 223–234, New York, NY, USA, 2006.

[11] H. Yoo, M. Shim, and D. Kim, "Self-synchronized and energy-aware duty cycle control algorithm in energy-harvested wireless sensor networks," *IEEE Communications Letters*, vol. 16, no. 2, pp. 203-204, 2012.

[12] L. Varga, G. Romaniello, and M. Favre, "GreenNet: an energy-harvesting IP-enabled wireless sensor network," *IEEE Internet of Things Journal*, vol. 2, no. 5, pp. 412–426, 2015.

[13] A. Kansal, J. Hsu, and S. Zahedi, "Power management in energy harvesting sensor networks," *ACM Transactions on Embedded Computing Systems*, vol. 6, no. 4, pp. 1–35, 2007.

[14] M. Z. Win, P. C. Pinto, and L. A. Shepp, "A mathematical theory of network interference and its applications," *Proceedings of the IEEE*, vol. 97, no. 2, pp. 205–230, 2009.

[15] A. Bletsas, A. G. Dimitriou, and J. N. Sahalos, "Improving backscatter radio tag efficiency," *IEEE Transaction on Microwave Theory and Techniques*, vol. 58, no. 6, pp. 1502–1509, 2010.

[16] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Transactions on Wireless Communications*, vol. 13, no. 1, pp. 418–428, 2014.

[17] S. J. Thomas, E. Wheeler, and J. Teizer, "Quadrature amplitude modulated backscatter in passive and semipassive UHF RFID systems," *IEEE Transaction on Microwave Theory and Techniques*, vol. 60, no. 4, pp. 1175–1182, 2012.

[18] G. Papotto, F. Carrara, A. Finocchiaro, and G. Palmisano, "A 90-nm CMOS 5-mbps crystal-less RF-powered transceiver for wireless sensor network nodes," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 2, pp. 335–346, 2014.

[19] C. Ok, H. Thadakamalla, and U. Raghavan, "Optimal transmission power in self-sustainable sensor networks for pipeline monitoring," in *2007 IEEE International Conference on Automation Science and Engineering*, Scottsdale, AZ, USA, 2007.

[20] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: a contemporary survey," *Networking and Internet Architecture*, 2014, http://arxiv.org/abs/1406.6470.

[21] L. Xia, J. Cheng, N. E. Glover, and P. Chiang, "0.56 V, −20 dBm RF-powered, multi-node wireless body area network system-on-a-chip with harvesting-efficiency tracking loop," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 6, pp. 1345–1355, 2014.

[22] S. Agrawal, S. K. Pandey, J. Singh, and M. S. Parihar, "Realization of efficient RF energy harvesting circuits employing different matching technique," in *2014 15th International Symposium on Quality Electronic Design (ISQED)*, pp. 754–761, Santa Clara, CA, USA, March, 2014.

[23] S. Leng, D. W. K. Ng, and R. Schober, "Power efficient and secure multiuser communication systems with wireless information and power transfer," in *2014 IEEE International Conference on Communications Workshops (ICC)*, Sydney, Australia, June, 2014.

[24] A. Seyedi and B. Sikda, "Energy efficient transmission strategies for body sensor networks with energy harvesting," *Transactions on Communications*, vol. 58, no. 7, pp. 2116–2126, 2010.

[25] H. P. Tan, P. W. Lee, W. K. Seah, and Z. A. Eu, "Impact of power control in wireless sensor networks powered by ambient energy harvesting (WSN-HEAP) for railroad health monitoring," in *WAINA '09. International Conference on Advanced Information Networking and Applications Workshops*, Bradford, UK, May 2009.

[26] D. M. Dobkin, *The RF in RFID: Passive UHF RFID in Practice*, Burlington, MA: Newnes, USA, 2008.

[27] T. S. Ajitha and V. Bhanumathi, "Improving RFID tag efficiency by QAM backscatter modulation," *International Journal of Scientific & Engineering Research*, vol. 4, no. 4, pp. 421–427, 2013.

[28] L. Burgess, "FCC and ETSI requirements for short-range UHF ASK modulated transmitters," *High Frequency Electronics*, vol. 12, pp. 28–35, 2005.

[29] R. Gonsalves, "Evaluation of outage probability in two-tier open access femtocell networks," in *2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, Mumbai, India, April 2014.

Hindawi
Journal of Sensors
Volume 2017, Article ID 9643279, 12 pages
https://doi.org/10.1155/2017/9643279

Research Article

A Model-Based Virtual Sensor for Condition Monitoring of Li-Ion Batteries in Cyber-Physical Vehicle Systems

Luciano Sánchez,[1] Inés Couso,[2] José Otero,[1] Yuviny Echevarría,[1] and David Anseán[3]

[1]*Computer Science Department, Universidad de Oviedo, Oviedo, Spain*
[2]*Statistics Department, Universidad de Oviedo, Oviedo, Spain*
[3]*Electrical Engineering Department, Universidad de Oviedo, Oviedo, Spain*

Correspondence should be addressed to Luciano Sánchez; luciano@uniovi.es

Received 7 March 2017; Accepted 27 August 2017; Published 12 October 2017

Academic Editor: Qing Tan

A model-based virtual sensor for assessing the health of rechargeable batteries for cyber-physical vehicle systems (CPVSs) is presented that can exploit coarse data streamed from on-vehicle sensors of current, voltage, and temperature. First-principle-based models are combined with knowledge acquired from data in a semiphysical arrangement. The dynamic behaviour of the battery is embodied in the parametric definition of a set of differential equations, and fuzzy knowledge bases are embedded as nonlinear blocks in these equations, providing a human understandable reading of the State of Health of the CPVS that can be easily integrated in the fleet through-life management.

1. Introduction

Cyber-physical vehicle systems (CPVSs) integrate locomotion, computational, and communication components, aiming to leverage interdependent behaviour by integrating control, computing, communications, and physical systems [1]. Monitoring, fault detection, and diagnosis of CPVSs are achieved through a combination of hardware sensors and decision-making software [2], often by "anytime" or "imprecise computing" algorithms that balance resources and performance, refining the solutions when the resources become available [3] or degrading gracefully with reduced cyber resources [4]. On the physical side, health monitoring of rechargeable batteries is an important part of both the monitoring and the energy management subsystems of a vehicle and measures the *battery ageing* that can manifest itself either as a gradually decreasing capacity (understood as the amount of electric charge that can be stored and released), a downtrending efficiency when the battery is charged or discharged or, in certain cases, as a catastrophic failure that destroys the cell [5].

Energy management algorithms for CPVS fleets should take into account the fact that the batteries in the vehicles are in different health conditions. Monitoring the battery health in due course has profound practical consequences, because there are silent deteriorations (those that cannot be perceived through a loss of capacity) that can trigger a sudden failure if not detected and acted upon. Ideally, the Battery Management Systems (BMSs) embedded in the powertrains of the vehicles should monitor the State of Health (SoH) of the batteries and notify the supervisor if a degradation is detected. However, as of yet there are no commercially available sensors of the health of a battery that can be used for this purpose. In this paper it is proposed that a virtual sensor (soft sensor) of the SoH is developed that combines signals already present in the BMS and also makes use of a battery model for synthesizing a "health signal" that is sent to the supervisor if an incipient degradation is detected.

Certain "first-principles" (electrochemical) models can estimate health-related variables taking current, voltage, and temperature as inputs [6]. Since these three inputs are available in standard BMSs, these models could, in principle, be part of the proposed virtual sensor. However, electrochemical processes during charge and discharge are different whether the battery is new or aged. Thus, first-principle models are not

effective in mutable scenarios, such as a fleet of CPVSs with different ages.

Learning models (equivalent circuits, statistical methods, neural networks, etc.) could be better suited for the problem being considered. It is remarked that pure data-driven models are not adequate either, because these models may generalize wrongly when subjected to unforeseen combinations of the input variables. For instance, a data-driven model that is learnt with data sampled from charge-discharge cycles at low currents will perform poorly when the current is high. Conversely, a "first-principles" model of a cell that is able to reproduce the behaviour of a new battery for both small and high currents will be inaccurate when applied to an ageing battery, as mentioned. A balance must be sought between the learning capabilities of the model and the amount of prior knowledge about the electrochemical processes that is placed in the model definition. To this we can add that the quality of on-vehicle measurements of current, voltage, and temperature is low, this being particularly true for the temperature. Generally speaking, the uncertainty in the values of the input variables lowers the accuracy of any kind of model, but some models are more resilient to uncertainty than others. Because of these reasons, a computational intelligence-based soft sensor is proposed that is intermediate between first-principle and pure data-driven models, as it is based on a "grey-box" model of the battery. By "grey-box" it is meant that the model is learnt from data, but at the same time it depends on physically meaningful parameters of the battery. The virtual sensor introduced in the present contribution is based on an "imprecise computing" model that exploits unreliable data streamed from on-vehicle sensors of current, voltage, and temperature. Parts of the proposed sensor are implemented with Fuzzy Rule-Based Systems (FRBS) that are fitted to operational data with Genetic Fuzzy Systems (GFSs).

The structure of the paper is as follows: in Section 2, the literature about fast methods for determining the health of a battery is reviewed. In Section 3, the proposed model is introduced. In Section 4, numerical results are provided. The paper concludes in Section 5 with some remarks and a list of yet-unsolved problems and challenges.

2. State of the Art in the State of Health Monitoring

The best reported methods for assessing battery health compute the functional dependence between the stored charge and the open circuit voltage (OCV) of the battery at equilibrium, by determining the positions of the peaks at the Incremental Capacity (ICA) [7]. This last curve is obtained by differentiating the battery charged capacity with respect to the OCV. In Figure 1(a), an example of the evolution in time of the ICA curve is shown, measured at one of the batteries used in this study. Alternatively, the inverse derivative (OCV with respect to the capacity) is the Differential Voltage (DVA) curve that also gives a clear insight of the efficiency of the battery (see Figure 1(b)). In both cases, a precise knowledge of the OCV is crucial for monitoring the condition of automotive batteries.

The most accurate method for obtaining the OCV is the "voltage relaxation" procedure [8]. Voltage relaxation consists in a sequence of short calibrated incremental charges (or discharges) that are combined to obtain a pointwise estimation of the OCV curve. However, this method is extremely slow and therefore unsuitable for on-board monitoring (in this particular case, the battery had to be removed from the vehicle for 4 days). In Figure 1(c), a graph showing voltage and current of a battery during an actual voltage relaxation essay is displayed.

Given that relaxation experiments are not a practical method for determining the OCV curve of the battery of a CPVS, accelerated methods are a must. However, these curves must be determined while the battery is at equilibrium, that is, when a charge (discharge) current is not flowing [9]. In practical circumstances, ICA and DVA analysis can be carried out when a current is flowing that is not higher than 1/25 of the capacity of the battery measured in Ah (or "C25" cycle); thus direct measurements are not reliable if they are completed in less than 25 hours [10].

Different procedures have been published where the OCV is approximated with data sampled during lapses shorter than a day, with varying accuracies; see, for instance, [11–15]. The problem is harder when the time window is shortened: if the battery has to be drained faster, the discharge current has to be raised accordingly and the memory effects of the battery are no longer negligible. Unfortunately, it is difficult to model the memory effects of a Li-Ion battery. As mentioned in Introduction, neither first-principles (or "white boxes") nor data-driven (or "black boxes") methods are valid. On the one hand, although the electrochemical, thermodynamic, and transport phenomena that define the behaviour of a Li-Ion battery are well known [16], white models are not valid because these models depend on a large number of parameters that are not provided by the manufacturer [17]. On the other hand, black boxes are unreliable when predicting unforeseen states and cannot incorporate prior knowledge about the mentioned phenomena [18]. A balance between white and black boxes is needed, and these are the "grey boxes," also known as "semiphysical models."

Grey models are a compromise where some parts of the definition of the model are taken from granted and other parts are learnt from data. The most prevalent grey boxes are equivalent circuit models, where the battery is assimilated to an electrical circuit comprising a network of resistances and capacitors [19]. SoH observers have been derived from equivalent circuit models; see, for instance, [20, 21]. These models produce estimations of the OCV whose accuracy is good (about ±1%) for a range of State of Charge (SoC) between 20 and 90%, provided that the charge/discharge current is between low and moderate. Since these equivalent circuit models are not based on electrochemical properties of the battery, the parameters of the OCV curve cannot be directly deduced from the values of the capacitors and resistances in the network. However, these models can be subjected to virtual experiments in accelerated time; that is, a relaxation experiment that lasts more than one day with real-world batteries can be simulated with the battery model in milliseconds. The "pseudo-OCV" curves that are

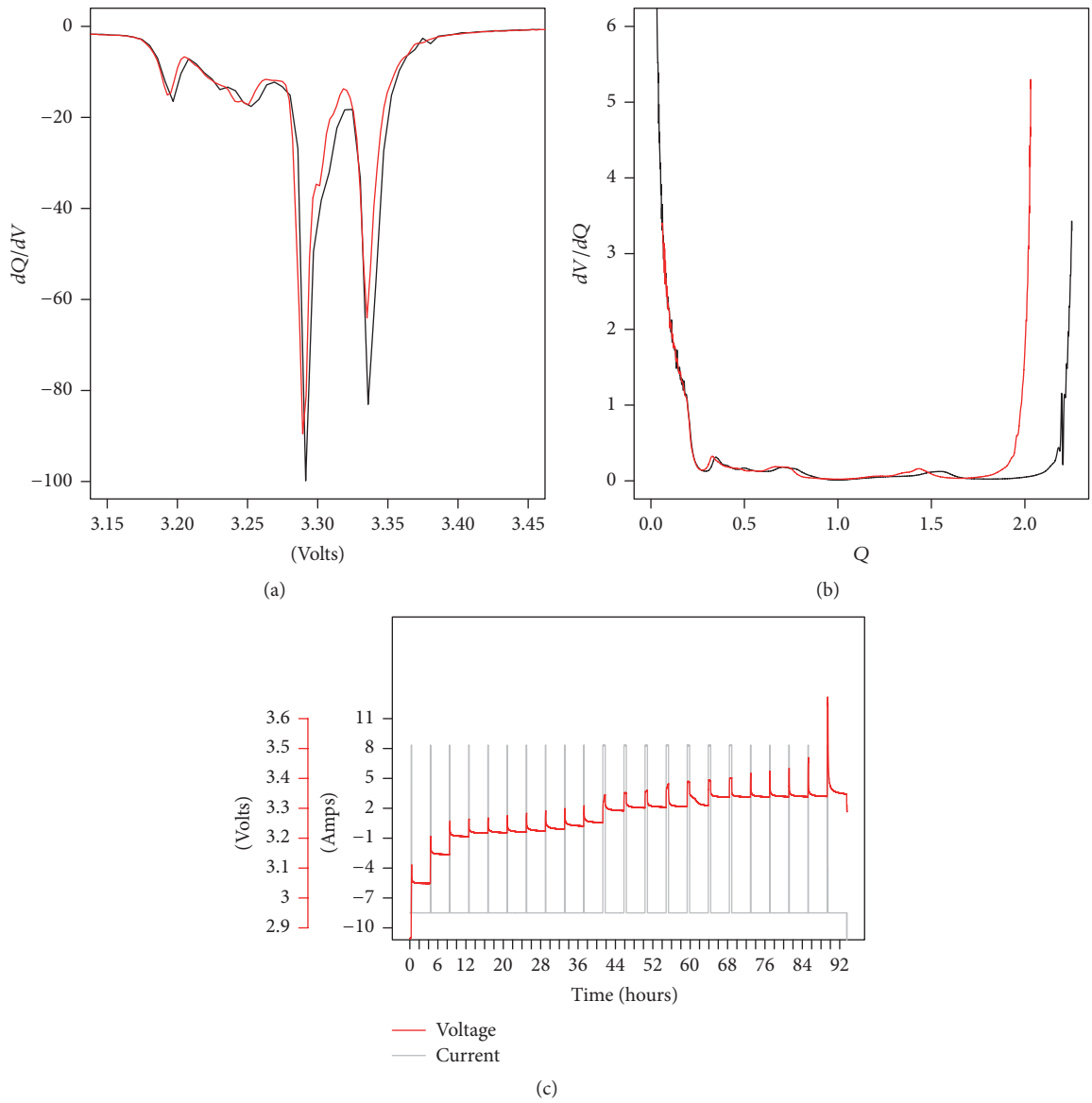


Figure 1: ICA (a) and DVA (b) curves of a new $LiFePO_4$ battery are drawn in black. These curves are superimposed in red for the same battery after 3000 charge-discharge cycles. The diagnostic of the battery is immediate when these curves are available. (c) Current and voltage during the experimental OCV determination by voltage relaxation of a typical $LiFePO_4$ battery. The relaxation time between partial charges is on 4 h. Pulses are of 10% of the battery capacity (42 Ah). OCV, IVA, and DVA can be measured at the laboratory, but the procedure is long, lasting more than four days in batteries with higher capacities. A soft sensor is sought that produces these curves through data sampled with on-vehicle sensors.

obtained from virtual experiments could, in principle, be processed to recover ICA or DVA graphs (see, e.g., [7]) but the accuracy of this kind of "virtual laboratory" experiments is poor.

3. A Semiphysical Model-Based Soft Sensor of the SoH

Other kinds of grey boxes have been successfully used in State of Charge (SoC) predictive models [22, 23], and these can be adapted to the problem at hand with certain modifications. These models are based on a physical analogy between charging a battery and filling a flexible vessel, because the function linking the height of a fluid with its mass, in a vessel with the appropriate shape, can also measure the interdependence between the voltage of a battery and its charge.

In this section, a design of an ensemble of Fuzzy Rule-Based Systems (FRBS) and differential equations is presented. The design follows the principles stated in [24] and is aimed to obtain an estimation of the SoH: after this ensemble is fitted to data, the KB of one of its member FRBS comprises a set of "if-then" rules describing the SoH of the battery through its OCV. This new definition allows that the SoH is inferred from the learnt parameters of the model without the need of a "virtual lab" relaxation experiment.

3.1. Notation. The on-vehicle signals are the charge current, $I(t)$, the battery voltage, $V(t)$, the battery temperature, $T(t)$, and the ambient temperature, $T_{\text{amb}}(t)$. The hidden variables are the electrical charge $\text{SoC}(t)$ and the overpotential $\text{OP}(t)$, which is the difference between the voltage of the battery and the OCV for the same SoC. The outputs of the soft sensor comprise the OCV curve of the battery (as a function of the SoC), $V(t)$, $T(t)$, and the hidden variables $\text{OP}(t)$ and $\text{SoC}(t)$, given the inputs $I(t)$, $T_{\text{amb}}(t)$ and the initial charge of the battery $\text{SoC}(0)$.

3.2. Battery Modelling with Fuzzy Rule-Based Systems. Batteries are complex systems. A black-box model is possible but it would require a long time window and training data covering many different scenarios, which is not always available (or attainable). However, the electrochemical processes that happen when batteries are charged are well known; thus there is a high potential for injecting expert knowledge into the model (by means of if-then rules, physical analogies, etc.) This addition of knowledge is intended to avoid that the algorithm that learns the model from data ends up confirming what is already known.

There are many different contributions regarding the balance between accuracy and interpretability in FRBS [25]. Most of these studies are intended to improve the linguistic interpretability of models that are intended for verbal communication (i.e., for explaining the model properties to a human expert). The proposed model conveys the reverse path, that is, injecting human knowledge; thus the model generalizes well for situations not present in the training data.

In this particular case, there are certain pieces of knowledge about the dynamic behaviour of the battery that can be efficiently represented by means of differential equations. An outline of the sensor structure is illustrated in Figure 2. The sensor comprises a combination of dynamical blocks including three FRBSs, representing the following.

(1) OCV versus SoC. This is the main output of the proposed soft sensor. The only input to this system is the SoC, which is the cumulative sum of the charge current, discounting the charging losses. There is a dotted input in the diagram, indicating that the OCV should also depend on the battery temperature, but this dependence can be safely ignored for LiFePO_4 batteries [26]. The output of this FRBS is the voltage of the battery at equilibrium.

(2) Overpotential in Steady State versus SoC and Temperature. This FRBS models the difference between the output voltage and the OCV when the battery is being charged (or discharged) at a constant pace. The inputs to this second system are the SoC and the battery temperature. Its output is not routed to the exterior but is fed to a feedback loop along with the net charging current that models the kinetic behaviour of the battery. The output of this feedback loop is added to the OCV to produce the second output of the sensor, the predicted battery voltage. The difference between the prediction of the battery voltage and the measured voltage is the first of the error signals which will be minimized during the learning process.

(3) Internal Calorific Power versus SoC and Current. The third FRBS models the heat emission of the battery as a function of the SoC and the charging current. The output of this system is not routed to the exterior either, but it is the input of an internal dynamical model of the temperature of the battery that depends on the specific heat of the cell and the thermal resistance between the cooling system and the ambient. The difference between the output of the dynamical model of the battery and the ambient temperature is the second error signal, which is also minimized during the learning.

3.3. Equations of the Sensor. The equations associated with the model in Figure 2 are detailed in this section. These equations encode the expert knowledge about the battery behaviour through the parametric definition of a set of differential equations. The available expertise about the battery dynamics is summarized in the following list.

(1) Nonreversible Energy Losses Are Proportional to the Charging Current, by a Factor That Depends on the Charge. Charging or discharging the battery is not a completely efficient process. There are energy losses and, to a lesser degree, charging losses that are handled by multiplying the input current by a factor (about 0.999 in this study; this is the "charging losses" box). Energy losses are modelled by a parasitic series voltage PV, proportional to the current ("power losses" box):

$$\text{Energy losses} = \left| I \cdot \text{PV}\left(\text{SoC}, I\right)\right|. \tag{1}$$

The absolute value is needed because the current is negative when discharging. In this model the simplification $\text{PV}(\text{SoC}, I) \approx k \cdot I$ is made, for a constant k that is learnt from data; thus the nonreversible energy losses are proportional to the square of the current.

(2) The Difference between the Voltage at Equilibrium and the Voltage While Charging (or Discharging) Depends on Charge, Current, and Temperature. The output voltage will be modelled as the sum of three terms: (a) the voltage in equilibrium, OCV, that depends on charge and temperature (this dependence is modelled by FRBS 1 in Figure 2), although in the particular case of LiFePO_4 technology the dependence between OCV and temperature was disregarded, as mentioned; (b) the overpotential, OP, that depends on charge, current, and temperature (FRBS 2 and diffusion process, in the same figure); (c) parasitic voltage associated with energy losses ("power losses" box)

$$V\left(\text{SoC}, T, I\right) = \text{FRBS}_1\left(\text{SoC}\right) + \text{OP}\left(\text{SoC}, T, I\right) + k \cdot I. \tag{2}$$

(3) The Voltage of the Battery in Open Circuit Keeps Changing for a Certain Time after a Charge or a Discharge Is Applied (Voltage Relaxation). Charging or discharging the battery involves diffusion processes, whose speed is limited: it is not physically possible to charge a battery in a very short time. According to our experimentation, a first-order differential equation or exponential decay with time constant τ is enough to acknowledge this constraint:

$$\tau \cdot \dot{\text{OP}} = -\text{OP} + I \cdot \text{FRBS}_2\left(\text{SoC}, T, \text{sign}\left(I\right)\right). \tag{3}$$

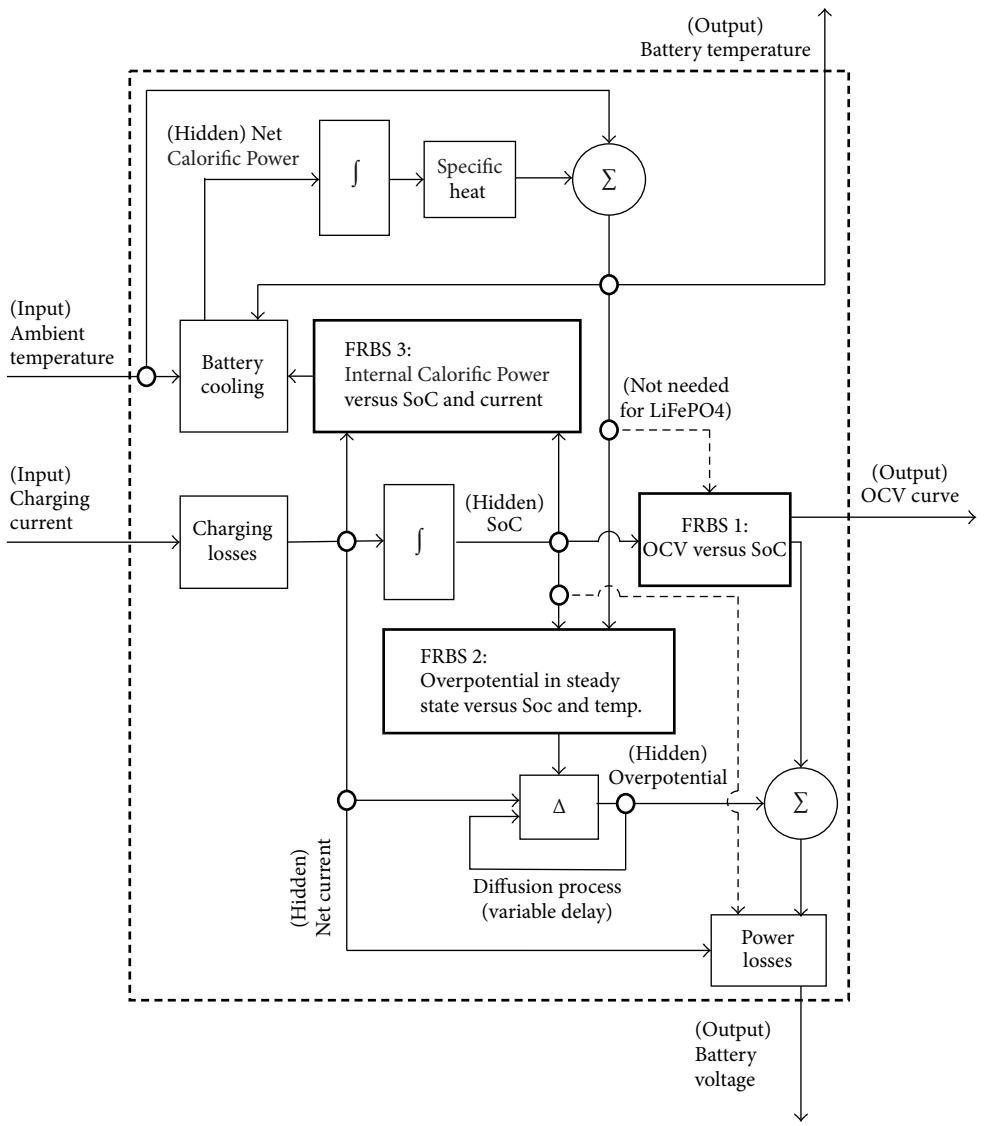


Figure 2: Block-based representation of the proposed soft sensor. OCV, overpotential, and Calorific Power are represented by FRBSs. These FRBSs can only be learnt from data in an indirect manner, by comparing the simulations of the model with the evolution of the observable variables and tuning the fuzzy rules in the three FRBSs until these trajectories become identical.

The forcing function is $I \cdot \mathrm{FRBS}_2(\mathrm{SoC}, T, \mathrm{sign}(I))$, where FRBS_2 can be understood as a charge-dependant impedance that models the linearized quotient between overpotential and current. If the current is 0, OP eventually also becomes 0, modelling the voltage relaxation mentioned before. Small batteries have low τ, that is, quick dynamics and vice versa.

(4) The Temperature of the Battery Depends on the Integral of the Difference between Generated and Dissipated Calorific Power. It is proposed that the dynamic properties of the battery temperature are approximated by this equation:

$$c \cdot \dot{T} = -\rho\left(T - T_{\mathrm{amb}}\right) + |I \cdot \mathrm{OP}| + I^2\left[k + T \cdot \mathrm{FRBS}_3\left(\mathrm{SoC}, \mathrm{sign}\left(I\right)\right)\right], \tag{4}$$

where c is the specific heat of the battery and ρ is the thermal conductance with the ambient. The term $I^2 k$ was explained in (1). The thermal power resulting from entropy changes, that is, $\partial(\mathrm{OCV})/\partial T$ in accordance with Helmholtz equation [27], is modelled by the triple product of (a) the output of FRBS_3; (b) the absolute temperature of the battery; and (c) the square of the current. It is remarked that the output of FRBS_3 may be negative, when the battery absorbs heat in a reversible process.

Summarizing, the differential equations describing the dynamics of the sensor model are

$$\dot{\mathrm{SoC}} = I$$

$$V = \mathrm{FRBS}_1 + \mathrm{OP} + k \cdot I$$

$$\tau \cdot \dot{\mathrm{OP}} = -\mathrm{OP} + I \cdot \mathrm{FRBS}_2$$

$$c \cdot \dot{T} = -\rho\left(T - T_{\text{amb}}\right) + |I \cdot \text{OP}| + I^2\left[k + T \cdot \text{FRBS}_3\right], \tag{5}$$

where $\text{FRBS}_1(\text{SoC})$, $\text{FRBS}_2(\text{SoC}, T, \text{sign}(I))$, and $\text{FRBS}_3(\text{SoC}, \text{sign}(I))$ are rule-based systems that are indirectly learnt from operation data, as described in the following subsection.

3.4. Learning Algorithm. Assuming that the battery temperature is controlled (thus there are no extreme temperature changes during the operation of the vehicle), the sensor equations can be discretized through the implicit Euler's method [28]:

$$\begin{aligned}
&\text{SoC}_{t+1} = \text{SoC}_t + \Delta t \cdot I_{t+1} \\
&\text{OP}_{t+1} = \frac{1}{\tau + \Delta t}\left[\tau \cdot \text{OP}_t + \Delta t \cdot I_{t+1}\right. \\
&\quad \left. \cdot\, \text{FRBS}_2\left(\text{SoC}_{t+1}, \text{sign}\left(I_{t+1}\right)\right)\right] \\
&V_{t+1} = \text{FRBS}_1\left(\text{SoC}_{t+1}\right) + \text{OP}_{t+1} + k \cdot I_{t+1} \\
&T_{t+1} \\
&\quad = \frac{\Delta t \cdot \left[\rho \cdot T_{\text{amb},t+1} + \left|I_{t+1} \cdot \text{OP}_{t+1}\right| + I_{t+1}^2 \cdot k\right] + c \cdot T_t}{c + \Delta t \cdot \left(\rho + I_{t+1}^2 \cdot \text{FRBS}_3\left(\text{SoC}_{t+1}, \text{sign}\left(I_{t+1}\right)\right)\right)},
\end{aligned} \tag{6}$$

where SoC_t stands for $\text{SoC}(t)$, SoC_{t+1} means $\text{SoC}(t + \Delta t)$, and so forth.

Assuming that SoC_0 is known and given a sequence of N samples of the input variables, I_t and $T_{\text{amb},t}$ for $t = 1, 2 \ldots, N$, then the computer simulation of the outputs V_t and T_t and the hidden variables OP_t and SoC_t consists in successive applications of (6), once for each time period t.

Given a sequence of on-vehicle sampled values V_t^*, T_t^* and the initial charge SoC_0^*, learning this model from data consists in determining

(1) the constants k, τ, ρ, and c,

(2) the rule-based systems $\text{FRBS}_1(\text{SoC})$, $\text{FRBS}_2(\text{SoC}, T, \text{sign}(I))$, and $\text{FRBS}_3(\text{SoC}, \text{sign}(I))$

that minimize the dissimilarities between the computer simulation and the measured values.

As mentioned, on-vehicle measurements are not reliable and nonlinear least squares are not robust in these conditions, because a single outlier can alter the results. The rule learning algorithm described in [29] is resilient to the presence of outliers and will be used here. This algorithm is based on the genetic optimization of a multivariate fuzzy-valued error which is a function of the parameters described before. This fuzzy function is described in the following paragraphs.

Let us suppose that the difference between the true temperature of the battery T_t^{true} and the perceived temperature T_t^{perc} is lower than a certain bound δ_α^T with a probability greater than $1 - \alpha$,

$$\begin{aligned}
\mathbf{T}_t^\alpha &= \left[T_t^{\text{perc}} - \delta_\alpha^T, T_t^{\text{perc}} + \delta_\alpha^T\right] \\
P\left(T_t^{\text{true}} \in \mathbf{T}_t^\alpha\right) &\geq 1 - \alpha.
\end{aligned} \tag{7}$$

Observe that the most specific estimation of the squared error of the model for each level α is

$$\text{err}_T^\alpha = \left\{\sum_{T=1}^{N}\left(T_t - \tau_t\right)^2 : \tau_t \in \mathbf{T}_t^\alpha\right\} \tag{8}$$

and the same course of reasoning can be applied to the voltage, for a different set of bounds δ_α^V:

$$\begin{aligned}
\mathbf{V}_t^\alpha &= \left[V_t^{\text{perc}} - \delta_\alpha^V, V_t^{\text{perc}} + \delta_\alpha^V\right] \\
\text{err}_V^\alpha &= \left\{\sum_{t=1}^{N}\left(V_t - \nu_t\right)^2 : \nu_t \in \mathbf{V}_t^\alpha\right\}.
\end{aligned} \tag{9}$$

Following [30], the nested families of sets err_T^α and err_V^α can be regarded as fuzzy sets $\widetilde{\text{err}}_T$ and $\widetilde{\text{err}}_V$, whose membership function is as follows:

$$\begin{aligned}
\widetilde{\text{err}}_T(\tau) &= \sup\left\{\alpha : \tau \in \text{err}_T^\alpha\right\} \\
\widetilde{\text{err}}_V(\nu) &= \sup\left\{\alpha : \nu \in \text{err}_V^\alpha\right\}
\end{aligned} \tag{10}$$

allowing the application of the genetic algorithm described in [29]. It is remarked that values δ_α^V and δ_α^T are tolerance intervals describing the accuracy of on-board sensors. If $\delta_\alpha^V = 0$ and $\delta_\alpha^T = 0$, the procedure described in this section reduces itself to ordinary nonlinear least squares; thus the estimation becomes precise if accurate sensors are available and degrades gracefully when quality information is not available.

4. Numerical Results

The experimental setup, comprising the battery type, the test equipment, and the charging and discharging protocols, is described in the first place. The numerical results of the experimentation are detailed in Section 4.2, including the compared results of the proposed method against a selection of fast OCV estimators.

4.1. Experimental Setup. The tested cell is a LiFePO_4 (LFP) pouch battery from European Batteries (see Figure 3). This cell uses a LFP cathode in combination with graphite anode active material. The rated capacity is 42 Ah at C/5 (the discharge current at C/5 is $42/5 = 8.4$ A). The average operating voltage is 3.2 V. The discharge and charge cut-off voltages are 2.5 V and 3.65 V, respectively. The dimensions in mm are $275 \times 166.5 \times 13.3$. The weight cell is 1010 g.

All tests have been done using a setup that consists of a SBT 10050 battery test system from PEC in combination with an ICP 750 climate chamber from Memmert. All tests were carried out at an ambient temperature of 23°C. Testing-machine adjustments were performed to improve the reliability and accuracy of the measurements. The first stage of the testing procedure was commissioning, during which the battery was identified and weighed. Then, a conditioning test was performed according to the USABC [8]. It consisted of three different constant current (C/3, C/2, and C) discharge cycles. The standard charging method

(a) (b)

Figure 3: (a) SBT 10050 battery test system from PEC and ICP 750 climate chamber. (b) Detail of the $LiFePO_4$ (LFP) pouch battery from European Batteries, rated at 42 Ah at C/5, used in the tests.

provided by the manufacturer was used to charge the cell. This consists of a constant current (CC) stage at C/2 until the cell reaches the charging cut-off voltage, followed by a constant voltage (CV) stage until the current decreases to 0.05 C. For subsequent testing, the battery capacity is considered stable when three successive C3/3 discharges agree within 2%.

After the capacity had stabilized, a full charge/discharge cycle was performed at a rate of C/25. The results of the C/25 measurements provide a practical capacity reference with minimal kinetic effects which is close to the maximum capacity attainable by the cell. The battery capacity was measured at C/5 and the result was 44.6 Ah instead of the rated capacity of 42 Ah. The measured capacity was taken into account for the open circuit voltage (OCV) at equilibrium measurements. To determine the battery OCV the voltage relaxation method was used [20]. With this method the battery voltage relaxes to the OCV at equilibrium after current interruption. To obtain the OCV at equilibrium at different states of charge (SoC) the voltage relaxation measurements were performed by charging and discharging a cell in 4460 mAh steps (approximately 10% of the measured capacity at C/5) at C/5 constant current. At the end of charge/discharge voltage, a constant voltage (CV) stage is applied until $I < \text{C}/100$ to ensure full charge/discharge voltage of the cell. Each charge and discharge step was followed by a rest period of 4 hours, after which the voltage was sampled. This voltage was assumed to be equal to the OCV at equilibrium.

Finally, the battery was subjected to discharges at C/5, C/3, C/2, and C constant currents. Discharging was carried out at CC until reaching the discharging cut-off voltage recommended by the manufacturer. There was an inactivity period of one hour after each charge or discharge until the battery temperature fell below 24.5°C.

4.2. Numerical Results and Discussion. In Figure 4 the OCV estimation produced by the proposed method is plotted in blue over the OCV points obtained with the relaxation method (in red). The hysteresis of the charge/discharge cycle used to determine the OCV is plotted along the data.

The slowest cycle at C/25 (50 hours) produced the most accurate estimation, which is in the same error range compared to the relaxation method for charges of 4 Ah and higher (SoC ≥ 10%). The accuracy of the estimation is reduced for C/5, C/3, C/2, and C, but the results are in the 20 mV range for charges as fast as C/2 (4 hours). Observe the excellent results when extrapolating the OCV to 45 Ah from all cycles faster than C/25. The determination of the OCV is not accurate for charges lower than 4 Ah (10% SoC); however this is a problem shared by all fast OCV estimators, as shown later.

The differences in accuracy as a function of the charge/discharge current are shown in detail in Figure 5. Observe that the relaxation method, being the most accurate experimental procedure itself, is also subjected to a small variance of ±10 mV, depending on whether the point was sampled when the cell is being charged or discharged. This variability has been displayed with the red error bars in Figure 5. In addition to this, the mean accuracy of the estimation is plotted against the experimental time in Figure 6, showing that the expected error grows quickly if the current is higher than C/2. The dotted superimposed grey curve is an exponential fitting to the data.

In Table 1 the compared error values (average of the absolute error in volts) of the proposed method and Xu et al.'s [21] and AbuSharkh and and Doerffel's [20] methods are displayed. Points of the relaxation OCV taken at charges higher than 40 Ah were discarded for the C/3, C/2, and C curves. A Friedman test was used to check that there are statistical differences between the methods. Observe that the proposed method was the best in all cases, with a single tie at C/25 between the proposed method and Xu et al.'s. Pairwise Mann–Whitney tests were also applied between the proposed method and each of the alternatives. The differences were regarded as significant when the p value of the Mann–Whitney test is lower than 0.05. Best results were marked in boldface.

To clearly perceive the differences between the proposed algorithm and the best alternative, the OCV estimation with the Xu et al.'s method is shown in Figure 7. Observe that the fitting is less accurate for charges below 8 Ah (20% SoC) or over 40 Ah (89% SoC), and the extrapolation to charges over the sampled data is not as regular as the soft sensor. The differences are clearer in Figure 8, showing that the alternatives are more efficient for small currents (the C/25 curve is not statistically different than the best) and the

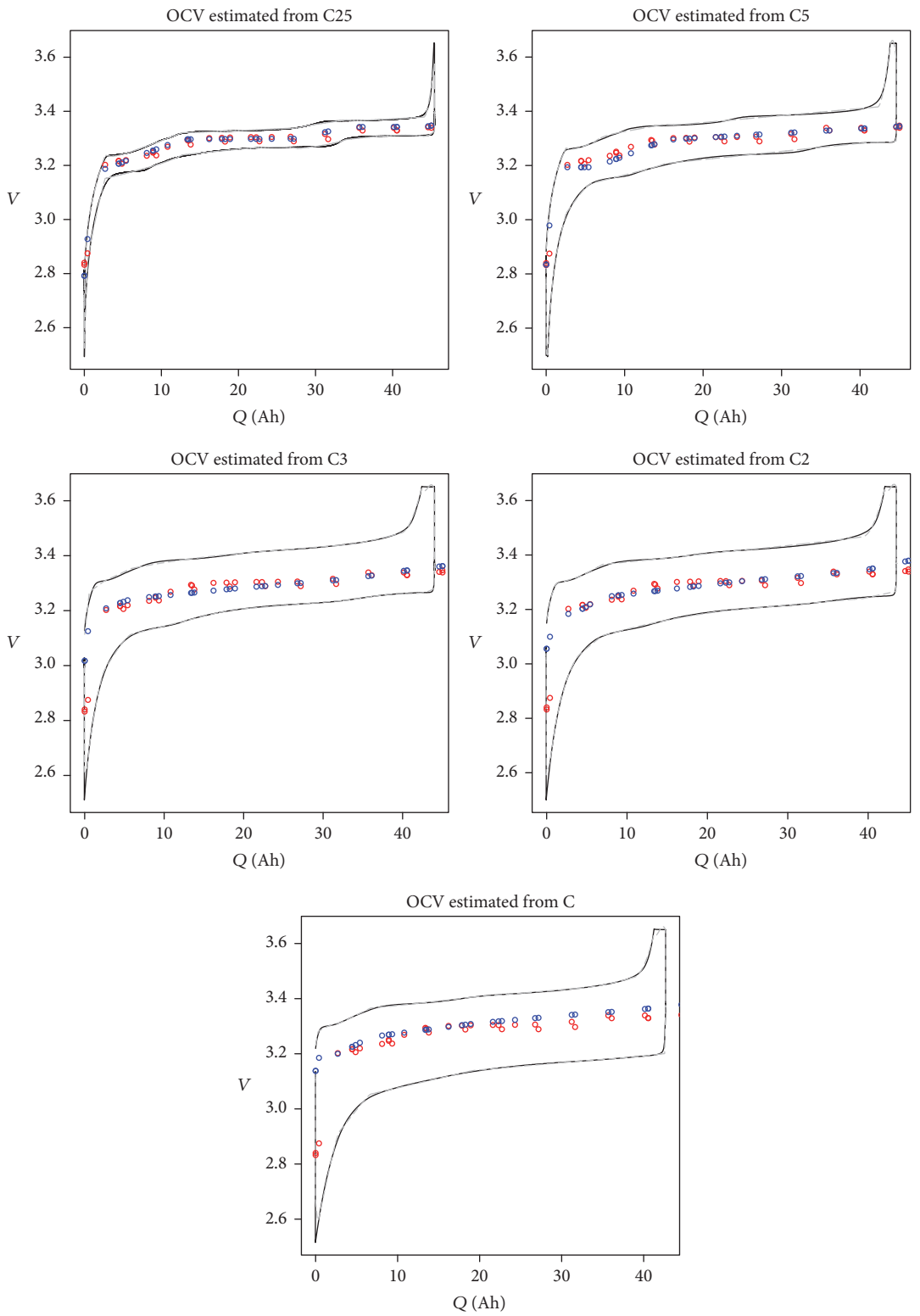


Figure 4: OCV estimated through the proposed method (in blue) over the OCV points obtained with the relaxation method (in red). The charge/discharge cycle used to learn the OCV is plotted along the data.

Table 1: Compared error values (average of the absolute error in volts) of the proposed method and Xu et al.'s [21] and AbuSharkh and Doerffel's [20] methods. The proposed method was the best in all cases, with statistically significant differences (p value of a Mann–Whitney test lower than 0.05). The only nonstatistical difference was at C/25 between the proposed method and Xu et al.'s.

Current	Proposed method	Upper limit (Ah)	Xu et al.'s method [21]	p value	AbuSharkh and Doerffel's method [20]	p value
C/25	**0.0094**	44	**0.0378**	0.1557	0.0393	0.0454
C/5	**0.0117**	44	0.0181	0.0046	0.0192	0.0007
C/3	**0.0117**	40	0.0181	0.0001	0.0192	$1e-05$
C/2	**0.0115**	40	0.0188	$1e-05$	0.0173	$6e-05$
C	**0.0186**	40	0.0188	$1e-05$	0.0188	$1e-05$

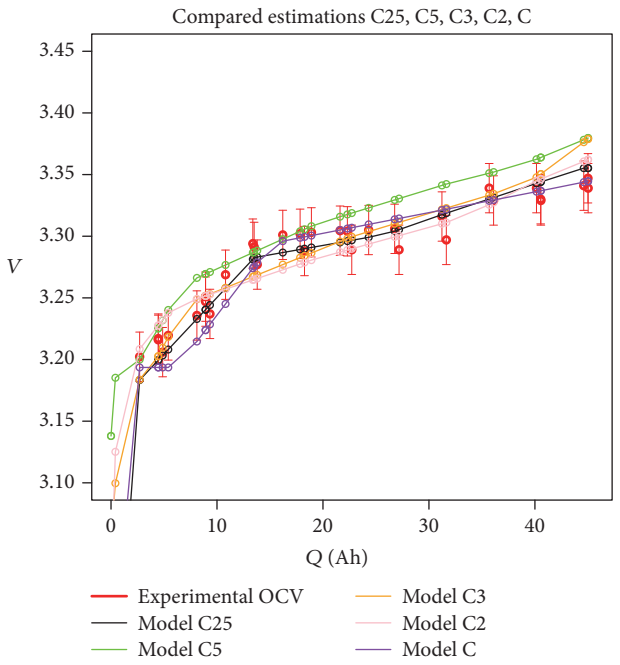


Figure 5: Compared results of the proposed method for charge currents of C/25, C/5, C/3, C/2, and C. The reference points obtained with the relaxation technique are plotted with red error bars.

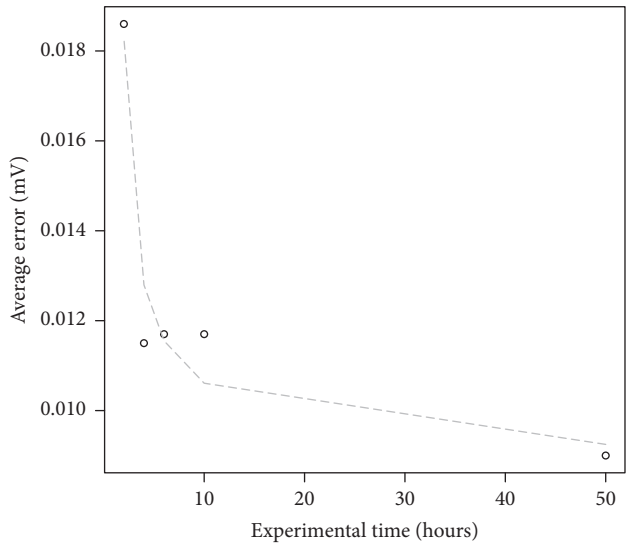


Figure 6: Mean accuracy of the estimation plotted against the experimental time. The expected error grows abruptly if the current is higher than C/2. The dotted grey curve is an exponential fitting to the data.

estimated OCV is not accurate for very low or very high charges, confirming the conclusions expressed by the authors of that method.

5. Concluding Remarks and Future Work

A novel model-based soft sensor for fault detection and diagnosis of rechargeable batteries for CPVSs has been proposed. The main contribution of the present work is the implementation of an "imprecise computing," learning semi-physical model of a battery. The proposed model contains three learnable FRBSs, connected with different dynamical blocks, in a setup that allows obtaining certain parameters of the underlying physical process that are costly or hard to estimate with a dedicated experiment. In particular, this virtual sensor is able to approximate the SoH of an automotive battery from on-vehicle measurements of current, voltage, and temperature, being resilient to inaccurate sensors. The present approximation is much faster than direct measurements by relaxation and the range of application of the new method extends that of the alternatives to SoCs under 20% and over 80%.

The results are promising but there is still a margin for improvement. The dependence between the time constants of the model, the current, and the charge is prone to overfitting, and it is possible that a new set of differential equations would exist that allows for a better fitting for SoCs below 10%. In future works, the first-order assumption for the dynamics of the overpotential will be dropped and replaced by a fractional-order differential equations model. Also, it is remarked that the experiments have been made in an off-vehicle temperature controlled environment. Further experiments are needed where the battery temperature is subjected to larger changes for different use patterns.

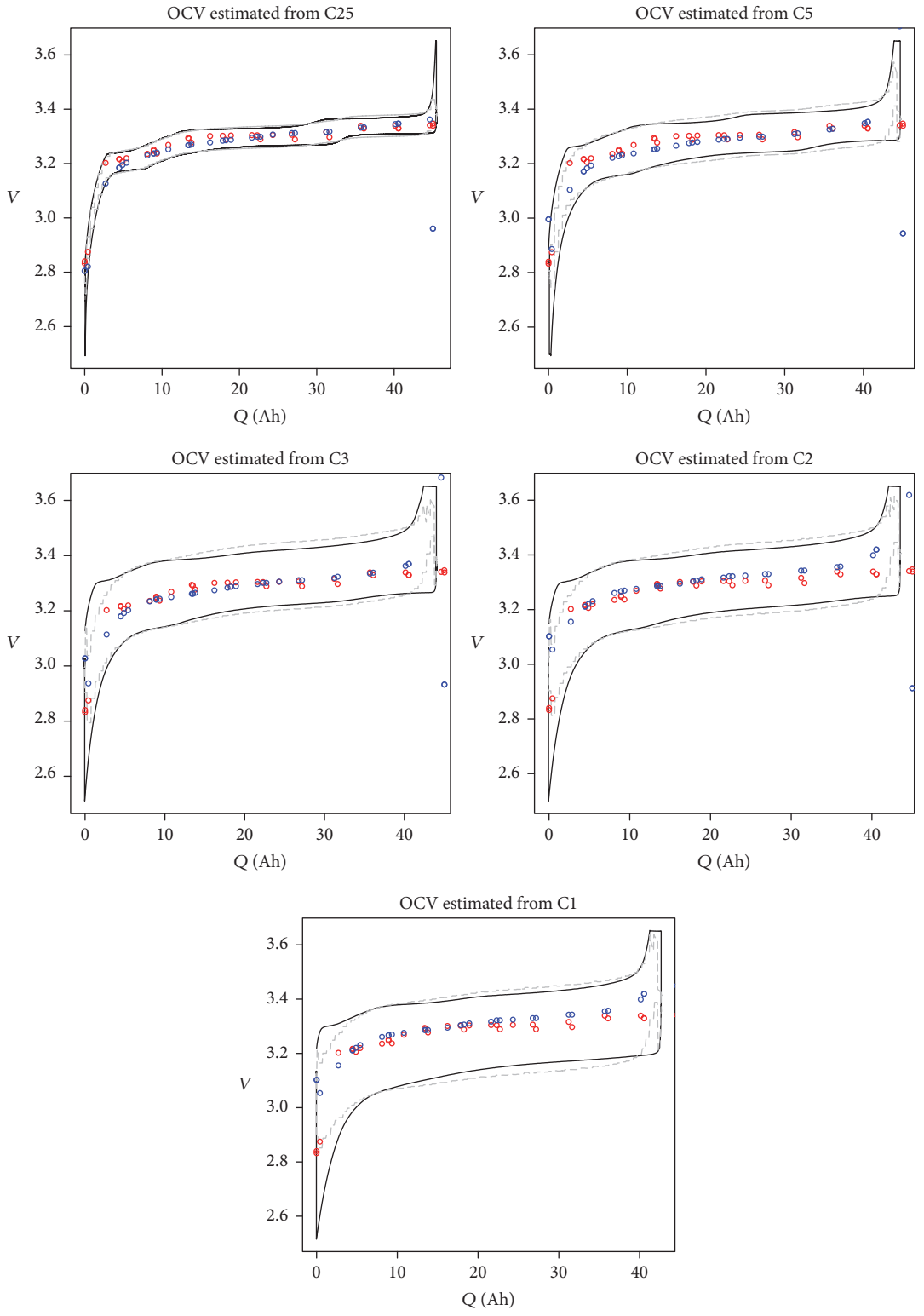


Figure 7: OCV estimated through Xu et al.'s method (in blue) over the OCV points obtained with the relaxation method (in red). The charge/discharge cycle used to learn the OCV is plotted along the data. Observe that the fitting is inaccurate for charges below 8 Ah (20% SoC) or over 40 Ah (89% SoC).

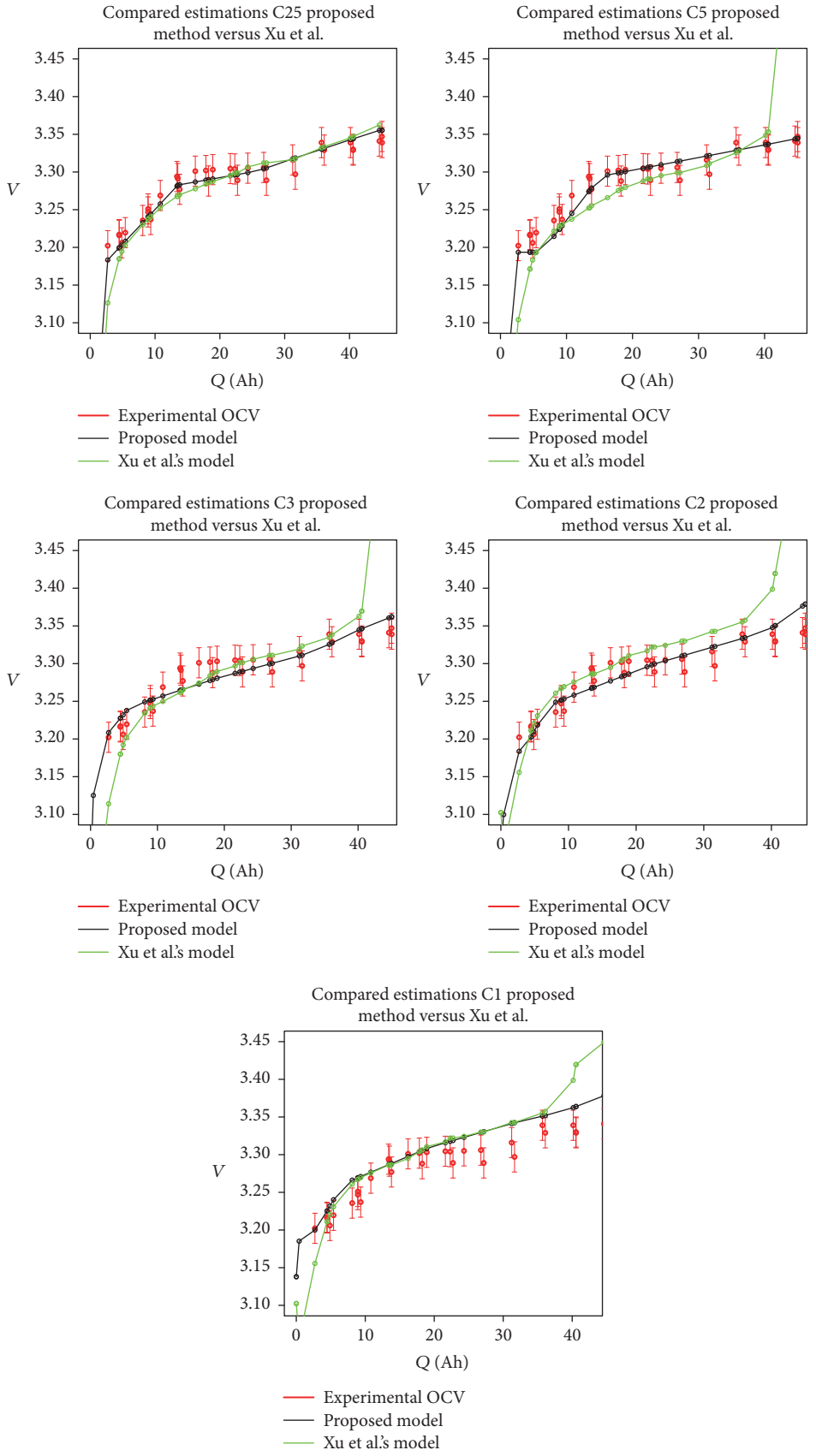


Figure 8: OCV estimations with a semiphysical fuzzy model and Xu et al.'s method. The fitting is less accurate for charges below 8 Ah (20% SoC) or over 40 Ah (89% SoC), and the extrapolation to charges over the sampled data is not as regular as semiphysical model.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Spanish Ministry of Science and Innovation (MICINN) and the Regional Ministry of the Principality of Asturias under Grants TIN2014-56967-R, TIN2017-84804-R, DPI2013-46541-R, and FC-15-GRUPIN14-073, in addition to the Eureka SD Project (Agreement no. 2013-2591) that is supported by the Erasmus Mundus Programme of the European Union.

References

[1] J. M. Bradley and E. M. Atkins, "Optimization and control of cyber-physical vehicle systems," *Sensors*, vol. 15, no. 9, pp. 23020–23049, 2015.

[2] N. Srivastava and J. Srivastava, "A hybrid-logic approach towards fault detection in complex cyber-physical systems," in *Proceedings of the Annual Conference of the Prognostics and Health Management Society*, vol. 1316, Portland, Oreg, USA, 2010.

[3] D. E. Quevedo and V. Gupta, "Sequence-based anytime control," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 377–390, 2013.

[4] S. J. Russell and S. Zilberstein, "Composing real-time systems," in *Proceedings of the 12th International Joint Conference on Artificial Intelligence (IJCAI '91)*, vol. 91, pp. 212–217, 1991.

[5] J. Vetter, P. Novák, M. R. Wagner et al., "Ageing mechanisms in lithium-ion batteries," *Journal of Power Sources*, vol. 147, no. 1-2, pp. 269–281, 2005.

[6] A. Barré, B. Deguilhem, S. Grolleau, M. Gérard, F. Suard, and D. Riu, "A review on lithium-ion battery ageing mechanisms and estimations for automotive applications," *Journal of Power Sources*, vol. 241, pp. 680–689, 2013.

[7] C. Weng, J. Sun, and H. Peng, "A unified open-circuit-voltage model of lithium-ion batteries for state-of-charge estimation and state-of-health monitoring," *Journal of Power Sources*, vol. 258, pp. 228–237, 2014.

[8] U.S.A.B. Consortium, "Electric vehicle battery test procedures manual," Tech. Rep., 1996.

[9] T. Reddy and D. Linden, *Linden's Handbook of Batteries*, McGraw-Hill, New York, NY, USA, 4th edition, 2010.

[10] D. Ansean, M. Gonzalez, V. Garcia, J. Viera, J. C. A. Anton, and C. Blanco-Viejo, "Evaluation of $LiFePO_4$ batteries for electric vehicle applications," *IEEE Transactions on Industry Applications*, vol. 51, no. 2, pp. 1855–1863, 2015.

[11] I. Snihir, W. Rey, E. Verbitskiy, A. Belfadhel-Ayeb, and P. H. L. Notten, "Battery open-circuit voltage estimation by a method of statistical analysis," *Journal of Power Sources*, vol. 159, no. 2, pp. 1484–1487, 2006.

[12] S. Tong, M. P. Klein, and J. W. Park, "On-line optimization of battery open circuit voltage for improved state-of-charge and state-of-health estimation," *Journal of Power Sources*, vol. 293, pp. 416–428, 2015.

[13] M. A. Roscher and D. U. Sauer, "Dynamic electric behavior and open-circuit-voltage modeling of $LiFePO_4$-based lithium ion secondary batteries," *Journal of Power Sources*, vol. 196, no. 1, pp. 331–336, 2011.

[14] T. Stockley, K. Thanapalan, M. Bowkett, and J. Williams, "Design and implementation of OCV prediction mechanism for PV-lithium ion battery system," in *Proceedings of the 20th International Conference on Automation & Computing (ICAC '14)*, pp. 49–54, September 2014.

[15] B. Xiao, Y. Shi, and L. He, "A universal state-of-charge algorithm for batteries," in *Proceedings of the 47th Design Automation Conference (DAC '10)*, p. 687, ACM Press, Anaheim, Calif, USA, June 2010.

[16] M. Yoshio, R. J. Brodd, and A. Kozawa, *Lithium-Ion Batteries*, Springer, New York, NY, USA, 2009.

[17] A. Seaman, T.-S. Dao, and J. McPhee, "A survey of mathematics-based equivalent-circuit and electrochemical battery models for hybrid and electric vehicle simulation," *Journal of Power Sources*, vol. 256, pp. 410–423, 2014.

[18] Y. Oussar and G. Dreyfus, "How to be a gray box: Dynamic semi-physical modeling," *Neural Networks*, vol. 14, no. 9, pp. 1161–1172, 2001.

[19] U. Westerhoff, K. Kurbach, F. Lienesch, and M. Kurrat, "Analysis of lithium-ion battery models based on electrochemical impedance spectroscopy," *Energy Technology*, vol. 4, no. 12, pp. 1620–1630, 2016.

[20] S. AbuSharkh and D. Doerffel, "Rapid test and non-linear model characterisation of solid-state lithium-ion batteries," *Journal of Power Sources*, vol. 130, no. 1-2, pp. 266–274, 2004.

[21] J. Xu, B. Cao, Z. Chen, and Z. Zou, "An online state of charge estimation method with reduced prior battery testing information," *International Journal of Electrical Power and Energy Systems*, vol. 63, pp. 178–184, 2014.

[22] C. Blanco, L. Sanchez, M. Gonzalez, J. C. Anton, V. Garcia, and J. C. Viera, "An equivalent circuit model with variable effective capacity for $LiFePO_4$ batteries," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 8, pp. 3592–3599, 2014.

[23] L. Sanchez, C. Blanco, J. C. Anton, V. Garcia, M. Gonzalez, and J. C. Viera, "A variable effective capacity model for $LiFePO_4$ traction batteries using computational intelligence techniques," *IEEE Transactions on Industrial Electronics*, vol. 62, no. 1, pp. 555–563, 2015.

[24] L. Sánchez, I. Couso, and M. González, "A design methodology for semi-physical fuzzy models applied to the dynamic characterization of $LiFePO_4$ batteries," *Applied Soft Computing*, vol. 14, pp. 269–288, 2014.

[25] J. Casillas, O. Cordón, F. H. Triguero, and L. Magdalena, *Interpretability Issues in Fuzzy Modeling*, vol. 128, Springer, 2013.

[26] J. Li, J. Klee Barillas, C. Guenther, and M. A. Danzer, "A comparative study of state of charge estimation algorithms for $LiFePO_4$ batteries used in electric vehicles," *Journal of Power Sources*, vol. 230, pp. 244–250, 2013.

[27] M. Hazewinkel, *Encyclopaedia of Mathematics: "Helmholtz Equation"*, Springer Science & Business Media, 2013.

[28] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, 2003.

[29] L. Sánchez, I. Couso, and J. Casillas, "Genetic learning of fuzzy rules based on low quality data," *Fuzzy Sets and Systems*, vol. 160, no. 17, pp. 2524–2552, 2009.

[30] I. Couso and L. Sánchez, "Higher order models for fuzzy random variables," *Fuzzy Sets and Systems*, vol. 159, no. 3, pp. 237–258, 2008.

Hindawi
Journal of Sensors
Volume 2017, Article ID 5757125, 14 pages
https://doi.org/10.1155/2017/5757125

Research Article

Quality-Aware Incentive Mechanism for Mobile Crowd Sensing

Ling-Yun Jiang,[1,2,3] Fan He,[1] Yu Wang,[1] Li-Juan Sun,[1,2,3] and Hai-ping Huang[1,2,3]

[1]*College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu 210003, China*
[2]*Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing, Jiangsu 210003, China*
[3]*Key Lab of Broadband Wireless Communication and Sensor Network Technology, Nanjing University of Posts and Telecommunications, Ministry of Education, Nanjing, Jiangsu 210003, China*

Correspondence should be addressed to Hai-ping Huang; hhp@njupt.edu.cn

Received 4 March 2017; Revised 24 July 2017; Accepted 22 August 2017; Published 28 September 2017

Academic Editor: Javier Sedano

Mobile crowd sensing (MCS) is a novel sensing paradigm which can sense human-centered daily activities and the surrounding environment. The impact of mobility and selfishness of participants on the data reliability cannot be ignored in most mobile crowd sensing systems. To address this issue, we present a universal system model based on the reverse auction framework and formulate the problem as the *Multiple Quality Multiple User Selection* (MQMUS) problem. The *quality-aware incentive mechanism* (QAIM) is proposed to meet the quality requirement of data reliability. We demonstrate that the proposed incentive mechanism achieves the properties of computational efficiency, individual rationality, and truthfulness. And meanwhile, we evaluate the performance and validate the theoretical properties of our incentive mechanism through extensive simulation experiments.

1. Introduction

A new paradigm of sensing with smartphones has emerged which is usually called people-centric mobile sensing or mobile crowd sensing [1]. Compared with the traditional sensor networks, MCS is an effective way for large-scale data sensing, processing, and gathering without deploying a large number of sensor nodes. MCS has enabled numerous large-scale applications such as urban environment monitoring [2–4], traffic flow surveillance [5–7], healthcare [8], behavior and relationship discovery [9, 10], indoor localization [11], 3G/Wi-Fi discovering [12–14], activity monitoring [15, 16], and bus arrival time prediction [17].

The effect of the aforementioned mobile crowd sensing applications relies heavily on the quantities of participants. However, the ordinary individuals are not willing to share their sensing capabilities unless there are sufficient incentives. Research on incentive mechanism has been widely concerned by investigators, and considerable designed schemes about the incentive mechanism design have been put forward which can be classified into nonmonetary incentives [18–20] and monetary incentives [21–29].

The key of any crowd sensing system is not only the quantities of participants but also the sensing quality offered by participants. However, most of the existing solutions usually assume that each sensing task (e.g., air quality in a certain region) in a sensing cycle could be performed by a single participant. It is intuitive that the quality of sensing project would be higher if each sensing task was performed by multiple participants. One of the main reasons is that the sensed data cannot always be trusted because participants maybe intentionally (e.g., malicious participants) or unintentionally (e.g., making mistakes) offer the data contrary to the truth. Another reason may come from the recruitment system model itself. A typical MCS consists of two roles: the recruiter who publicizes the sensing tasks and the participants who constitute potential sensing capability selected by the recruiter from many candidates. The interaction between the recruiter and the candidates is modeled as a reverse auction in many existing solutions which can be illustrated by Figure 1. The recruiter always selects participants according to the sensing plans of the candidates. However, changes always go beyond plans. The participants may not be able to complete the task according to their schedule for unexpected incidents

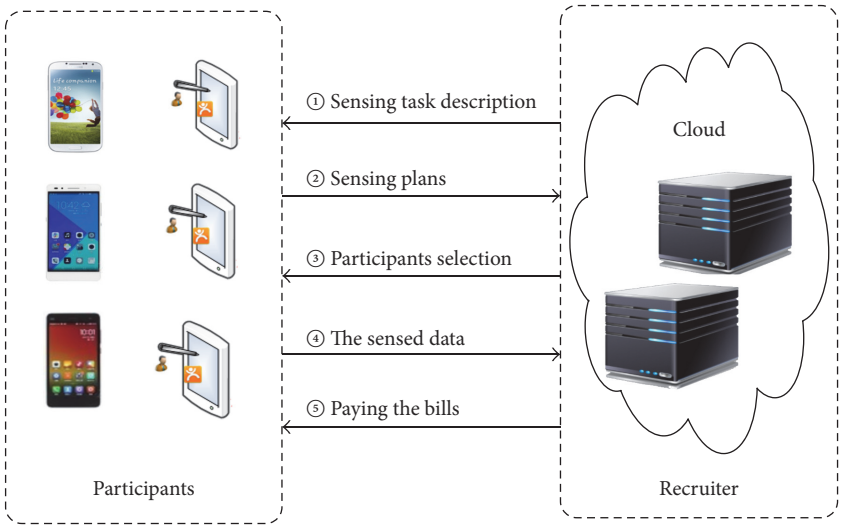


Figure 1: A typical mobile crowd sensing system as a reverse auction framework.

(e.g., a selected participant cannot go to the specific locations claimed in his sensing plan). These participants may offer some forged data or do nothing. As a result, the tasks could not be completed in time.

In this paper, we address the issue of quality-aware monetary incentive mechanism design. We design a truthful incentive mechanism satisfying the properties of computational efficiency, individual rationality, and truthfulness with low approximation ratio.

The remainder of this paper is organized as follows. In Section 2, we review the related work. In Section 3, we describe the system model and formulate the MQMUS problem. Thereafter, in Section 4 we propose the incentive mechanism, named QAIM, which consists of two phases, winner selection and payment determination, and analyze the properties of QAIM. Section 5 presents the experimental results. Finally, we draw the conclusion and discuss some possible future directions in Section 6.

2. Related Work

There are lots of incentive mechanisms which can be classified into nonmonetary incentives [18–29] and monetary incentives [30–45]. Paying for sensed data in crowd sensing tasks is the most intuitive incentive. Monetary incentive mechanisms are mainly based on two kinds of schemes: Stackelberg game and auction.

Stackelberg game is a game where one leader player has the dominant influence over the other players [46]. Duan et al. [30] make use of the Stackelberg game to design a threshold revenue model for service providers. The system and the users interact through a two-stage process similar to that of Stackelberg game. The system announces the total reward and the threshold number of required participants. Each participant decides whether to accept the task or not. Yang et al. [31] also model the proposed platform-centric incentive mechanism as a Stackelberg game, prove that this Stackelberg game has a unique equilibrium, and design an efficient mechanism for computing it. The above two Stackelberg game solutions have theoretical guarantees. However, the premise of this kind of method is that the costs of all users or their probability distributions are assumed to be known, which limits the applicability of Stackelberg game-based mechanisms because participants may keep their costs private in the real world.

An auction-based mechanism is originally the process of buying and selling goods by negotiating the monetary prices [47]. A kind of auction, called reverse auction, is adopted to model the negotiation process in crowd sensing, which is shown in Figure 1. Lee and Hoh [32] firstly design a reverse auction-based dynamic price incentive mechanism with virtual participation credit with the objective of minimizing and stabilizing the platform cost while maintaining the participation level. Yang et al. [31] consider two system models for smartphone crowd sensing system: the platform-centric model with the solution based on the Stackelberg game and user-centric model with the solution based on the reverse auction. Feng et al. [33] formulate the winning bids determination problem and present a truthful auction for location-aware collaborative sensing. Zhang et al. [34] focus on the user-centric model and study three methods which involve cooperation and competition among the services. Xu et al. [35, 36] investigate truthful incentive mechanisms for time window dependent tasks with the strong requirement of data integrity and propose two incentive mechanisms for the single time window case and the multiple window case, respectively. Subramanian et al. [37] consider offline and online incentive mechanisms using the same bidding framework with MSensing Auction proposed in [31]. Zhao et al. [38] investigate the incentive mechanisms in the online setting based on an offline budget feasible mechanism [39], which provides a starting point for the online mechanism. Jin et al. [40] pay attention to the quality of the mobile crowd sensing systems and incorporate a metric named QoI

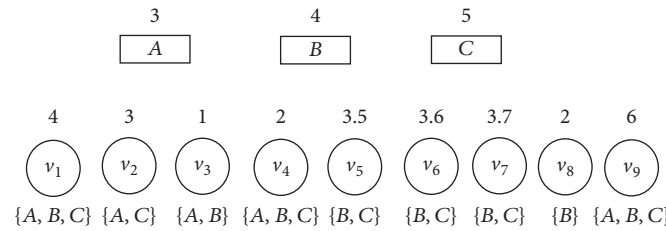


Figure 2: A motivating example with diversity requirements.

(Quality of Information) into the incentive mechanisms. SRC and MRC mechanisms with the criterion of the combinatorial QoI and price are proposed. However, the authors fail to consider the truthfulness of the MRC mechanism. The aforementioned solutions assume that each measurement of sensing task can be represented by a single sensor reading.

Several solutions are proposed to ensure the quality of crowd sensing data. Tanas and Herrera-Joancomartí [48] achieve the first work, which focuses on how to validate sensing data, but the premise of their work is that there are multiple users to submit multiple sensing readings on each task. Kazemi et al. [49] assume each worker has a reputation score, and assign enough number of workers to each spatial task such that workers' aggregate reputation can satisfy the confidence of the task. However, they focus on self-incentivized spatial crowdsourcing, in which people perform the tasks voluntarily without any reward. Zhang et al. [41] propose a task management framework to match workers to the merged query and sensing tasks efficiently. In their model, each task can be assigned to multiple workers, and each worker can be assigned to at most one task, although each worker may have the preference for multiple tasks. Xu et al. [42] design the incentive mechanism, which considers the issue of stimulating the biased requesters in the competing crowdsourcing market. Xiong et al. [43] consider the k-depth coverage as an MCS data collection constraint, but every subtask is assigned to the same value of k. Wang et al. [44] present a detailed quality-aware mobile crowdsourced sensing framework, composed of three MCS components: crowd, crowdsourcer, and crowdsourcing platform. The crowdsourcer is a new role who assesses the posted contributions' quality. He et al. [45] propose a recruitment strategy in vehicle-based crowdsourcing through taking full advantage of predictable mobility patterns of vehicles, which bring a new insight to improve the quality of crowd sensing system. However, the behaviors of human are affected by many factors. It is far more difficult to predict the mobility patterns of human beings than those of vehicles.

In this paper, we try to enhance the quality-aware incentive mechanism from two main dimensions: the reputation of participants and the design of task.

3. Problem Statement

Different from most crowd sensing systems, the objective of this paper is designing the truthful incentive mechanism with maximum social efficiency and high sensing quality. To achieve this objective, the recruiter needs to select participants who can match the diverse requirements of the crowd sensing application with minimum social cost. Before demonstrating the rigorous problem definition, we would like to present a motivating example to make the problem better understood.

3.1. A Participant Recruitment Example in Air Quality Monitoring. We take the urban air quality monitoring MCS task as an example. As shown in Figure 2, the MCS recruiter wants to collect the state of the air in three regions (denoted as $G = \{A, B, C\}$). Nine candidates ($\{v_1, v_2, \ldots, v_9\}$) are interested in performing the task and reporting their sensing plans, which include what they can do with the corresponding bid price. The industrial structures vary greatly in different regions. The regions with more plants, which can discharge waste gas, need more participants to monitor. For example, the recruiter wants 5 participants to monitor region C and only 3 to monitor region A because there are more chemical plants in region C. We use squares to represent the regions, and the number above each square denotes its requirement. To the perspective of the candidates, people may not just stay in a certain region in one sensing cycle and can fulfill multiple sensing tasks in different regions. We use disks to represent the candidates, and the number above each disk denotes its corresponding bid price, and the set of regions below each disk denotes the regions that he can monitor.

In this example, the mobile crowd sensing system has some requirements: (1) Every subtask should be assigned to enough participants so that their aggregate sensing results can ensure the sensing quality. (2) Every subtask has different sensing requirement. The different number of participants should be recruited to satisfy different sensing requirements with minimum costs. (3) Every participant has different ability in terms of the task completion and should be assigned to the different number of subtasks based on his particular ability.

3.2. System Model and Problem Formulation. We present the rigorous definition and formulation of the MQMUS problem. In this problem, the recruiter can divide the task into multiple subtasks with different quality factors and the participants can be assigned to multiple subtasks in one sensing cycle.

Suppose that a crowd sensing task G can be divided into e disjoint subtasks according to the sensing geographic areas,

and each subtask g_k has its sensing quality factor $h(g_k)$ (to simplify, we use the number of participants to represent $h(g_k)$ as shown in the above motivating example). The recruiter publicizes the sensing task $G = \{g_1, g_2, \ldots, g_k, \ldots, g_e\}$ and the quality factor $h(G) = \{h(g_1), h(g_2), \ldots, h(g_k), \ldots, h(g_e)\}$ as a quality constraint for participants selecting.

Considering n candidates, $U = \{v_1, v_2, \ldots, v_i, \ldots, v_n\}$ are interested in performing the sensing task. Each candidate $v_i \in U$ submits a sensing plan $B_{v_i} = (\psi_{v_i}, b_{v_i})$ to the recruiter, in which $\psi_{v_i} = \{g_1^{v_i}, g_2^{v_i}, \ldots, g_k^{v_i}\}$ is the set of subtasks that candidate v_i can perform (the superscript v_i of $g_k^{v_i}$ is only used to represent that v_i can fulfill the subtask g_k) and b_{v_i} is bid price that candidate v_i wants to charge for performing ψ_{v_i}.

We assume that the candidate v_i has a reputation score r_{v_i}, which states the probability that the candidate performs a task correctly. The recruiter is responsible for maintaining and updating the reputation score of every candidate. The value of r_{v_i} is set to 1 initially and updated by

$$r_{v_i} = r_{v_i} + \sum_{g_m \in \psi_{v_i}} \frac{(h(g_m) - 1)\eta}{\max_{g_k \in G}\{h(g_k)\}}. \tag{1}$$

We utilize a voting mechanism to set the value of η. This intuition is based on the idea of the wisdom of crowds [50] that the majority of the participants are trusted. The recruiter aggregates the different sensing results to get the reliable result at the end of the sensing cycle. The setting way of η is inspired by [44]. η is set to "−1" in two cases: (1) the candidate cannot perform the subtask as the claimed sensing plan; (2) the sensing result of the same subtask is contrary to more than half of participants' results; otherwise, η is "0." If $r_{v_i} < 0$, v_i will not be selected until the recruiter resets r_{v_i} to 1 after a period of time (e.g., 10 sensing cycles).

Assume that the number of candidates is sufficient to fulfill the sensing task G with its quality constraints $h(G)$. This assumption is reasonable for mobile crowd sensing systems as made in [31, 33, 35]. The selected participant s_j is placed into the list $S = \{s_1, s_2, \ldots, s_{j-1}\}$ according to the order. s_j is the ID of the candidate and its subscript j denotes that s_j is the jth selected participant. The recruiter has to calculate the payment p_{s_j} for each participant as the incentive. The utility of participant can be calculated by (2), in which c_{s_j} is the real cost of the participant s_j and only known by itself. b_{s_j} is not less than c_{s_j} due to the selfishness and rationality of participants (if the reputation score of s_j is set to a value less than 0 in this sensing cycle, the utility of s_j will be 0 in the next sensing cycle because he will not be selected).

$$u_{s_j} = p_{s_j} - c_{s_j}. \tag{2}$$

The utility of the recruiter is calculated by (3). $V(h(G))$ is the value to the recruiter when it has collected enough data to satisfy the quality constraints $h(G)$ of the sensing task G.

$$u_0 = V(h(G)) - \sum_{s_j \in S} p_{s_j}. \tag{3}$$

The social efficiency of the sensing task G (with the quality constraints $h(G)$) is calculated by (4). Although the real cost c_{s_j} is only known by participant s_j, we will prove that claiming a different cost b_{s_j} cannot help to increase the utility of participant s_j in our designed mechanisms. So we use b_{s_j} when we attempt to maximize social efficiency in the mechanisms designed below. The objective of maximizing the social efficiency is equivalent to the objective of minimizing the social cost.

$$u_{h(G)} = V(h(G)) - \sum_{s_j \in S} b_{s_j}. \tag{4}$$

Given the list of selected participants $S = \{s_1, s_2, \ldots, s_i, \ldots, s_m\}$, G_{s_i} is the set of the remaining subtasks excluding those subtasks of participants $\{s_1, s_2, \ldots, s_i\}$ according to their sensing plans. The goal of achieving high quality crowd sensing with minimum social cost can be formulated as (5) and constrained by (6).

$$\min \quad \left(\sum_{s_j \in S} b_{s_j}\right) \tag{5}$$

$$\text{s.t.} \quad \left|\psi_{s_1} \cap G\right| + \left|\psi_{s_2} \cap G_{s_1}\right| + \cdots + \left|\psi_{s_m} \cap G_{s_{m-1}}\right| \geq \sum_{g_k \in G} h(g_k). \tag{6}$$

We design a truthful incentive mechanism, QAIM, to select appropriate participants to satisfy the objective of this paper, and to eliminate the fear of market manipulation (the participants cannot improve their utility by submitting a bid price different from its real cost).

QAIM consists of two phases: winner selection algorithm QAIM(S) and payment determination algorithm QAIM(P). For a given $h(G)$ and a set of bids $B = \{B_{v_1}, B_{v_2}, \ldots, B_{v_i}, \ldots, B_{v_n}\}$, the algorithm QAIM($S$) selects a subset of participants $S \subseteq U$ and the algorithm QAIM(P) returns the vector $(p_1, p_2, \ldots, p_m)$ for those selected participants.

We cannot find the optimal solution in polynomial time for the MQMUS problem presented in (5) and (6) because this problem is NP-hard. The proof is in Appendix.

Our objective is to design the incentive mechanisms satisfying the following four desirable properties to solve MQMUS problem:

(i) *Computational Efficiency*. A mechanism is computationally efficient if both the winner selection function and payment decision function can be computed in polynomial time.

(ii) *Individual Rationality*. Each participant will have a nonnegative utility upon performing the sensing task.

(iii) *Truthfulness*. A mechanism is truthful if no participant can improve its utility by submitting a bid price different from its real cost, no matter what others submit. In other words, reporting the real cost is a dominant strategy for all participants.

(iv) *Social Optimization*. The objective function is maximizing the social efficiency. We attempt to find optimal solution or approximation algorithm with low approximation ratio when there is no optimal solution computed in polynomial time.

Input: $h(G)$, set of bids B
(1) $\lambda = 1$; $G_\lambda = G$; $S = \Phi$; for $l \in U$ if $(r_l < 0)$ $U = U - \{l\}$;
(2) while $(G_\lambda \neq \Phi)$ {
(3) for each $l \in (U - S)$ if $(\varpi_l(G_\lambda) \neq 0)$ $T_l(\lambda) = \dfrac{b_l}{\varpi_l(G_\lambda)}$ else $T_l(\lambda) = \mathrm{MAX}$;
(4) for each $l \in (U - S)$ Sort $T_l(\lambda)$ in non-decreasing order;
(5) $q = 1$; $y = 1$; $\min h_\lambda = \min_{g_k \in G_\lambda}\{h(g_k)\}$;
(6) while $(y \leq \min h_\lambda)$ {
(7) $i = \arg(T_q(\lambda))$; $S = S \cup \{i\}$; $y = y + 1$; $q = q + 1$;
(8) for each $(g_k^i \in \psi_i)$ {
if $(g_k^i \in G_\lambda)$ {
(9) $h(g_k) = h(g_k) - 1$;
(10) if $h(g_k) = 0$ $G_{\lambda+1} = G_\lambda - \{g_k\}$; } } }
(11) $\lambda = \lambda + 1$; }
(12) return S

Algorithm 1: QAIM(S).

input: $h(G)$, set of bids B, list of selected participants S
(1) for all $u \in U$ do $\{p_u = 0;\}$ for $l \in U$ if $(r_l < 0)$ $U = U - \{l\}$;
(2) for all $s_i \in S$ do {
(3) $U' = U - \{s_i\}$;$\chi = \Phi$; $\upsilon = 1$; $G_\upsilon = G$;
(4) while $(G_\upsilon \neq \Phi)$ {
(5) for $l \in (U' - \chi)$ {if $(\varpi_l'(G_\upsilon) \neq 0)$ $\Gamma_l(\upsilon) = \dfrac{b_l}{\varpi_l'(G_\upsilon)}$ else $\Gamma_l(\upsilon) = \mathrm{MAX};$}
(6) for $l \in (U' - \chi)$ {Sort $\Gamma_l(\upsilon)$ in non-decreasing order;}
(7) $q = 1$; $y = 1$; $\min h = \min_{g_k \in G_\upsilon}\{h(g_k)\}$;
(8) while $(y \leq \min h)$ {
(9) $r = \arg(\Gamma_q(\upsilon))$; $p_{s_i} = \max\{p_{s_i}, \dfrac{\varpi_{s_i}'(G_\upsilon)}{\varpi_r'(G_\upsilon)} b_r\}$;
$\chi = \chi \cup \{r\}$; $y = y + 1$; $q = q + 1$;
(10) for $(g_k^r \in \psi_r)$ {
(11) if $(g_k^r \in G_\upsilon)$ {
(12) $h(g_k) = h(g_k) - 1$;
(13) if $h(g_k) = 0$ $\{G_{\upsilon+1} = G_\upsilon - \{g_k\};$ } } } }
(14) $\upsilon = \upsilon + 1$; } }
(15) return P

Algorithm 2: QAIM(P).

4. Mechanism Design and Analysis

4.1. Mechanism Design. We attempt to find an approximation algorithm following a greedy approach which can be solved in polynomial time because the MQMUS problem is NP-hard problem. The winner selection algorithm QAIM(S) is illustrated in Algorithm 1 and the payment algorithm QAIM(P) is illustrated in Algorithm 2.

In Algorithm 1, l is the ID of candidates, λ is the number of selection round, and G_λ is the set of remaining subtasks excluding those in the sensing plans of the selected participants before the previous $\lambda - 1$ rounds. The *effective sensing units* of l in the λth round are denoted by $\varpi_l(G_\lambda)$ which can be calculated by (7), the *effective average sensing weight* of candidate l in the λth round is denoted by $T_l(\lambda)$ which is calculated in Line (3) of Algorithm 1.

$$\varpi_l(G_\lambda) = \left|G_\lambda \cap \psi_l\right|. \tag{7}$$

The main idea of greedy approach is to select candidate with least effective average sensing weight, so $T_l(\lambda)$ of all remaining candidates are sorted in nondecreasing order in Line (4) of Algorithm 1, and $\arg(T_q(\lambda))$ is the ID of the qth selected participant in the λth selection round.

The trick of QAIM(S) lies in the use of $\min h_\lambda$ which denotes the number of participants that can be selected in the

λth round. The nondecreasing sorting of $T_l(\lambda)$ implies that (8) is true.

$$\frac{b_{\arg(T_1(\lambda))}}{\varpi_{\arg(T_1(\lambda))}(G_\lambda)} \le \frac{b_{\arg(T_2(\lambda))}}{\varpi_{\arg(T_2(\lambda))}(G_\lambda)} \le \cdots \le \frac{b_{\arg(T_{\min h_\lambda}(\lambda))}}{\varpi_{\arg(T_{\min h_\lambda}(\lambda))}(G_\lambda)}. \tag{8}$$

Equation (9) is true; otherwise the first selected participant in the $(\lambda+1)$th round will be selected in the λth round.

$$\frac{b_{\arg(T_{\min h_\lambda}(\lambda))}}{\varpi_{\arg(T_{\min h_\lambda}(\lambda))}(G_\lambda)} \le \frac{b_{\arg(T_1(\lambda+1))}}{\varpi_{\arg(T_1(\lambda+1))}(G_\lambda)}. \tag{9}$$

The calculation method of $G_{\lambda+1}$ in Line (10) of Algorithm 1 implies that $|G_{\lambda+1}|$ cannot be bigger than $|G_\lambda|$, so (10) is true which implies the participant is selected in the nondecreasing order of the *effective average sensing weight*:

$$\frac{b_{\arg(T_{\min h_\lambda}(\lambda))}}{\varpi_{\arg(T_{\min h_\lambda}(\lambda))}(G_\lambda)} \le \frac{b_{\arg(T_1(\lambda+1))}}{\varpi_{\arg(T_1(\lambda+1))}(G_\lambda)} \le \frac{b_{\arg(T_1(\lambda+1))}}{\varpi_{\arg(T_1(\lambda+1))}(G_{\lambda+1})}. \tag{10}$$

Let $s_p, s_{p+1}, \ldots, s_{p+\min h_\lambda - 1}$ denote the IDs of the selected participants in the λth selection round; the set of remaining subtasks would possibly be changed only at the end of the λth selection round, so (11) is true.

$$G_{s_p} = G_{s_{p+1}} = \cdots = G_{s_{p+\min h_\lambda - 2}} = G_\lambda, \quad G_{s_{p+\min h_\lambda - 1}} = G_{\lambda+1}. \tag{11}$$

Equation (12) is true by derivation from (10) and (11).

$$\frac{b_{s_1}}{\varpi_{s_1}(G)} \le \frac{b_{s_2}}{\varpi_{s_2}\left(G_{s_1}\right)} \le \cdots \le \frac{b_{s_i}}{\varpi_{s_i}\left(G_{s_{(i-1)}}\right)} \le \cdots \le \frac{b_{s_m}}{\varpi_{s_m}\left(G_{s_{(m-1)}}\right)}. \tag{12}$$

In Algorithm 2, u is the ID of candidates with the same role as l in QAIM(S), υ is the number of payment determination round, and G_υ is the set of remaining subtasks excluding those in the sensing plans of the selected participants before the previous $\upsilon-1$ rounds. The *effective sensing units* of l in the υth round are denoted by $\varpi'_l(G_\upsilon)$ with the same calculation method used in (7). The *effective average sensing weight* of candidate l in the υth round is denoted by $\Gamma_l(\upsilon)$ which is calculated in Line (5) of Algorithm 2.

To compute the payment for each s_i in the winner list S, we consider the set of candidates $U - \{s_i\}$ and reselect appropriate participants into the list χ^{s_i} with the same method used in QAIM(S) (the superscript s_i of χ^{s_i} is used to identify that s_i is not considered as a candidate). Let $\chi^{s_i} = \{x_1^{s_i}, x_2^{s_i}, \ldots, x_k^{s_i}, \ldots, x_a^{s_i}\}$ and $x_k^{s_i}$ denote the kth selected participant, $G_{x_k^{s_i}}$ is the set of remaining subtasks excluding those effective subtasks of participants $\{x_1^{s_i}, x_2^{s_i}, \ldots, x_k^{s_i}\}$ according to their sensing plans, and (13) is true for the same reason of (12).

$$\frac{b_{x_1^{s_i}}}{\varpi'_{x_1^{s_i}}(G)} \le \frac{b_{x_2^{s_i}}}{\varpi'_{x_2^{s_i}}\left(G_{x_1^{s_i}}\right)} \le \cdots \le \frac{b_{x_k^{s_i}}}{\varpi'_{x_k^{s_i}}\left(G_{x_{k-1}^{s_i}}\right)} \le \cdots \le \frac{b_{x_a^{s_i}}}{\varpi'_{x_a^{s_i}}\left(G_{x_{a-1}^{s_i}}\right)}. \tag{13}$$

4.2. A Walk-Through Example. To better understand the algorithm, we use the example in Figure 2 to illustrate how the QAIM works.

With regard to the aforementioned example, the crowd sensing task $G = \{A, B, C\}$ is divided into 3 subtasks: $h(A)$ is set to 3, $h(B)$ is set to 4, and $h(C)$ is set to 5. There are 9 candidates $U = \{v_1, v_2, \ldots, v_9\}$ who want to participate in the task and report their sensing plan: the bid price of v_i is shown above it, and the subtask that v_i can fulfill is given below it in Figure 2. Take v_3, for example, $\psi_{v_3} = \{A, B\}$, which can also be represented as $\psi_{v_3} = \{A^{v_3}, B^{v_3}\}$, and $b_{v_3} = 1$. The *effective sensing units* of v_3 in the first round are denoted by $\varpi_{v_3}(G_1)$ which is calculated by (7) (i.e., $|\{A, B, C\} \cap \{A, B\}| = 2$), and the *effective average sensing weight* of v_3 in the first round is denoted by $T_{v_3}(1)$ which is calculated in Line (3) of Algorithm 1 (i.e., $b_{v_3}/\varpi_{v_3}(G_1) = 1/2$).

We first assume that all participants are trustworthy and can fulfill the sensing units as they had claimed in their sensing plan.

In the first selection round, $G_\lambda = \{A, B, C\}$, $h(A) = 3$, $h(B) = 4$, $h(C) = 5$, $\min h_\lambda = 3$. $T_l(\lambda)$ of each candidate in the first round is listed in Table 1. According to QAIM(S), v_3 is the first winner and then v_4, and v_1 is the third one in the first round. The selected list $S = \{v_3, v_4, v_1\}$, $G_{v_3} = \{A, B, C\}$, $G_{v_4} = \{A, B, C\}$, and $G_{v_1} = \{B, C\}$.

In the second selection round, $G_\lambda = \{B, C\}$, $h(A) = 0$, $h(B) = 1$, $h(C) = 3$, and $\min h_\lambda = 1$. $T_l(\lambda)$ of each candidate in the second round is listed in Table 2. According to QAIM(S), v_5 is the first and only winner. The selected list $S = \{v_3, v_4, v_1, v_5\}$ and $G_{v_5} = \{C\}$.

In the third selection round, $G_\lambda = \{C\}$, $h(A) = 0$, $h(B) = 0$, $h(C) = 2$, and $\min h_\lambda = 2$. $T_l(\lambda)$ of each candidate in the third round is listed in Table 3. According to QAIM(S), v_2 is the first winner and v_6 is the second one. The selected list $S = \{v_3, v_4, v_1, v_5, v_2, v_6\}$, $G_{v_2} = \{C\}$, and $G_{v_6} = \varnothing$.

In the fourth selection round, $G_\lambda = \varnothing$, which implies the selected participants $(v_3, v_4, v_1, v_5, v_2, v_6)$ can satisfy the sensing requirements.

If v_3 is a malicious participant, he lies in the results of all sensing units; the reputation of v_3 is 0 calculated by (1) which means he will not be selected in the next sensing recruitment cycle.

Owning to the limitation of the space and the similarity of the algorithm process, we only give the payment example

Table 1: $T_l(\lambda)$ in the first selection round.

	$T_{v_1}(\lambda)$	$T_{v_2}(\lambda)$	$T_{v_3}(\lambda)$	$T_{v_4}(\lambda)$	$T_{v_5}(\lambda)$	$T_{v_6}(\lambda)$	$T_{v_7}(\lambda)$	$T_{v_8}(\lambda)$	$T_{v_9}(\lambda)$
$\lambda = 1$	4/3	3/2	1/2	2/3	3.5/2	3.6/2	3.7/2	2/1	6/3

Table 2: $T_l(\lambda)$ in the second selection round.

	$T_{v_2}(\lambda)$	$T_{v_5}(\lambda)$	$T_{v_6}(\lambda)$	$T_{v_7}(\lambda)$	$T_{v_8}(\lambda)$	$T_{v_9}(\lambda)$
$\lambda = 2$	3/1	3.5/2	3.6/2	3.7/2	2/1	6/2

Table 3: $T_l(\lambda)$ in the third selection round.

	$T_{v_2}(\lambda)$	$T_{v_6}(\lambda)$	$T_{v_7}(\lambda)$	$T_{v_8}(\lambda)$	$T_{v_9}(\lambda)$
$\lambda = 3$	3/1	3.6/1	3.7/1	MAX	6/1

Table 4: $\Gamma_l(\upsilon)$ in the first payment determination round considering v_3.

	$\Gamma_{v_1}(\upsilon)$	$\Gamma_{v_2}(\upsilon)$	$\Gamma_{v_4}(\upsilon)$	$\Gamma_{v_5}(\upsilon)$	$\Gamma_{v_6}(\upsilon)$	$\Gamma_{v_7}(\upsilon)$	$\Gamma_{v_8}(\upsilon)$	$\Gamma_{v_9}(\upsilon)$
$\upsilon = 1$	4/3	3/2	2/3	3.5/2	3.6/2	3.7/2	2/1	6/3

Table 5: $\Gamma_l(\upsilon)$ in the second payment determination round considering v_3.

	$\Gamma_{v_5}(\upsilon)$	$\Gamma_{v_6}(\upsilon)$	$\Gamma_{v_7}(\upsilon)$	$\Gamma_{v_8}(\upsilon)$	$\Gamma_{v_9}(\upsilon)$
$\upsilon = 2$	3.5/2	3.6/2	3.7/2	2/1	6/2

Table 6: $\Gamma_l(\upsilon)$ in the third payment determination round considering v_3.

	$\Gamma_{v_6}(\upsilon)$	$\Gamma_{v_7}(\upsilon)$	$\Gamma_{v_8}(\upsilon)$	$\Gamma_{v_9}(\upsilon)$
$\upsilon = 3$	3.6/1	3.7/1	MAX	6/1

of the first selected winner v_3 which is similar to the payment determination of other participants. p_{v_3} is initialized to 0.

In the first payment determination round, $G_\upsilon = \{A, B, C\}$, $h(A) = 3$, $h(B) = 4$, $h(C) = 5$, and $\min h_\upsilon = 3$. $\Gamma_l(\upsilon)$ of each candidate in the first round is listed in Table 4. According to QAIM(P), v_4 is the first winner; then $p_{v_3} = \max\{p_{v_3}, (\varpi'_{v_3}(G_\upsilon)/\varpi'_{v_4}(G_\upsilon))b_{v_4}\} = \max\{0, (2/3) * 2\} = 4/3$. v_1 is the second winner; then $p_{v_3} = \max\{p_{v_3}, (\varpi'_{v_3}(G_\upsilon)/\varpi'_{v_1}(G_\upsilon))b_{v_1}\} = \max\{4/3, (2/3) * 4\} = 8/3$. v_2 is the third winner; then $p_{v_3} = \max\{p_{v_3}, (\varpi'_{v_3}(G_\upsilon)/\varpi'_{v_2}(G_\upsilon))b_{v_2}\} = \max\{8/3, (2/2) * 3\} = 3$.

In the second payment determination round, $G_\upsilon = \{B, C\}$, $h(A) = 0$, $h(B) = 1$, $h(C) = 2$, and $\min h_\upsilon = 1$. $\Gamma_l(\upsilon)$ of each candidate in the second round is listed in Table 5. According to QAIM(P), v_5 is the first and only winner; then $p_{v_3} = \max\{p_{v_3}, (\varpi'_{v_3}(G_\upsilon)/\varpi'_{v_5}(G_\upsilon))b_{v_5}\} = \max\{3, (1/2) * 3.5\} = 3$.

In the third payment determination round, $G_\upsilon = \{C\}$, $h(A) = 0$, $h(B) = 0$, $h(C) = 1$, and $\min h_\upsilon = 1$. $\Gamma_l(\upsilon)$ of each candidate in the second round is listed in Table 6. According to QAIM(P), v_6 is the first and only winner; then $p_{v_3} = \max\{p_{v_3}, (\varpi'_{v_3}(G_\upsilon)/\varpi'_{v_6}(G_\upsilon))b_{v_6}\} = \max\{3, (0/1) * 3.5\} = 3$.

In the fourth payment round, $G_\upsilon = \varnothing$, so the payment to v_3 is 3.

4.3. Properties of QAIM. In this section, we analyze the properties of QAIM theoretically to show that QAIM is computationally efficient, individually rational, and truthful. The approximation is also discussed in the end. We use n to denote the number of candidates, e to denote the number of subtasks, $\Omega(G)$ to denote all the sensing units, and $\Omega(G'_{s_i})$ to denote the effective sensing units of the selected participants $S = \{s_1, s_2, \ldots, s_i\}$, and for ease of analysis they can be calculated by

$$\Omega(G) = \sum_{k=1}^{e} h(g_k), \tag{14}$$

$$\Omega\left(G'_{s_i}\right) = \left|\psi_{s_1} \cap G\right| + \left|\psi_{s_2} \cap G_{s_1}\right| + \cdots + \left|\psi_{s_i} \cap G_{s_{i-1}}\right|. \tag{15}$$

(1) QAIM Is Computationally Efficient. We analyze QAIM(S) and QAIM(P), respectively, where QAIM takes $O(n^3 * e)$ in the worst case.

The nested for-loop (Lines (8)–(11)) of QAIM(S) will be executed $\lambda * |\psi_{s_i}| * \min h_\lambda$ times. The maximal value of λ is $\Omega(G)$ which is less than n obviously in the worst case when the *effective sensing unit* of every candidate has only one subtask. $|\psi_{s_i}|$ is less than e and $\min h_\lambda$ is far less than n obviously, so QAIM(S) takes $O(n^2 * e)$ in the worst case.

QAIM(P) takes $O(n^3 * e)$ in the worst case because there are similar processes in both QAIM(S) (Lines (2)–(11)) and QAIM(P) (Lines (4)–(14)), which will be executed $|S|$ times less than n.

(2) QAIM Is Individually Rational. When considering the set of candidates $U - \{s_i\}$, let $x_k^{s_i}$ be the replacement of participant s_i which appears in the kth place in the selected list $\chi^{s_i} =$

$\{x_1^{s_i}, x_2^{s_i}, \ldots, x_k^{s_i}, \ldots, x_a^{s_i}\}$. Equation (16) is true according to the main idea of winner selection.

$$s_1 = x_1^{s_i},\ s_2 = x_2^{s_i}, \ldots, s_{i-1} = x_{i-1}^{s_i}, \quad \left(\text{That is } G_{s_{i-1}} = G_{x_{i-1}^{s_i}}\right). \tag{16}$$

$x_k^{s_i}$ will not be selected in the kth place if s_i is considered, so (17) is true.

$$\frac{b_{s_i}}{\varpi_{s_i}\left(G_{s_{i-1}}\right)} \le \frac{b_{x_i^{s_i}}}{\varpi_{x_i^{s_i}}\left(G_{s_{i-1}}\right)}, \quad \left(\text{That is } \frac{b_{s_i}}{G_{s_{i-1}} \cap \psi_{s_i}} \le \frac{b_{x_i^{s_i}}}{G_{s_{i-1}} \cap \psi_{x_i^{s_i}}}\right). \tag{17}$$

Equation (18) is true based on the derivation from (16) and (17).

$$\frac{b_{s_i}}{G_{x_{i-1}^{s_i}} \cap \psi_{s_i}} = \frac{b_{s_i}}{G_{s_{i-1}} \cap \psi_{s_i}} \le \frac{b_{x_i^{s_i}}}{G_{s_{i-1}} \cap \psi_{x_i^{s_i}}} = \frac{b_{x_i^{s_i}}}{G_{x_{i-1}^{s_i}} \cap \psi_{x_i^{s_i}}}, \quad \left(\text{That is } b_{s_i} \le \frac{G_{x_{i-1}^{s_i}} \cap \psi_{s_i}}{G_{x_{i-1}^{s_i}} \cap \psi_{x_i^{s_i}}} b_{x_i^{s_i}}\right). \tag{18}$$

Equation (19) is true according to the main idea of payment calculation in Line (9) of QAIM(P).

$$p_{s_i} = \max_{x_r^{s_i} \in \chi^{s_i}} \left\{ \frac{\varpi'_{s_i}\left(G_{x_{r-1}^{s_i}}\right)}{\varpi'_{x_r^{s_i}}\left(G_{x_{r-1}^{s_i}}\right)} b_{x_r^{s_i}} \right\}. \tag{19}$$

From the analysis of (18) and (19), we know $b_{s_i} \le p_{s_i}$.

(3) QAIM Is Truthful. According to Myerson's Theorem [51], an auction is truthful if and only if the selection rule is monotone and each winner is paid the critical value p_i; if a participant wins the auction by bidding b_i, he also wins by bidding $b'_i < b_i$ but loses by bidding $b'_i > p_i$.

The monotonicity of the selection rule is obvious: if s_i bids a smaller b'_{s_i} that means $b'_{s_i}/\varpi_{s_i}(G_{s_{(i-1)}}) \le b_{s_i}/\varpi_{s_i}(G_{s_{(i-1)}})$, s_i will also be selected according to (12).

Suppose $p_{s_i} = \max_{x_r^{s_i} \in \chi^{s_i}}\{(\varpi'_{s_i}(G_{x_{r-1}^{s_i}})/\varpi'_{x_r^{s_i}}(G_{x_{r-1}^{s_i}}))b_{x_r^{s_i}}\} = (\varpi'_{s_i}(G_{x_{f-1}^{s_i}})/\varpi'_{x_f^{s_i}}(G_{x_{f-1}^{s_i}}))b_{x_f^{s_i}}$; (20) is true if b_{s_i} is greater than p_{s_i}.

$$b_{s_i} \ge \frac{\varpi'_{s_i}\left(G_{x_{f-1}^{s_i}}\right)}{\varpi'_{x_f^{s_i}}\left(G_{x_{f-1}^{s_i}}\right)} b_{x_f^{s_i}}, \quad \left(\text{That is } \frac{b_{s_i}}{\varpi'_{s_i}\left(G_{x_{f-1}^{s_i}}\right)} \ge \frac{b_{x_f^{s_i}}}{\varpi'_{x_f^{s_i}}\left(G_{x_{f-1}^{s_i}}\right)}\right). \tag{20}$$

Equation (20) shows the fact that s_i will not be selected before the previous f participants $(x_1^{s_i}, x_2^{s_i}, \ldots, x_f^{s_i})$ are selected. But if the previous f participants are selected, there is no reason to select s_i because the previous selected participants can satisfy different sensing requirements.

(4) The Approximation Factor to Optimal Solution Is $\ln(\Omega(G)) + 1$. Let OPT denote the minimal social cost computed by optimal solution, $\Omega(G'_{s_i})$ denote the effective sensing units of the selected participants $S = \{s_1, s_2, \ldots, s_i\}$ calculated by (15), and $\text{cost}(s_{i+1})$ denote the social cost of the $(i+1)$th selected participant s_{i+1}.

Because the participant is selected in the nondecreasing order of the *effective average sensing weight* according to QAIM(S) and the average cost of the rest uncovered sensing units is not greater than $\text{OPT}/(\Omega(G) - \Omega(G'_{s_i}))$, (21) is true.

$$\text{cost}\left(s_{i+1}\right) \le \frac{\text{OPT}}{\Omega\left(G\right) - \Omega\left(G'_{s_i}\right)}. \tag{21}$$

Hence the total cost of QAIM can be calculated by

$$\begin{aligned} \sum_{r=1}^{m} \text{cost}\left(s_r\right) &\le \sum_{r=1}^{m} \frac{\text{OPT}}{\Omega\left(G\right) - \Omega\left(G'_{s_{r-1}}\right)} \\ &\le \frac{\text{OPT}}{\Omega\left(G\right)} + \frac{\text{OPT}}{\Omega\left(G\right) - 1} + \frac{\text{OPT}}{\Omega\left(G\right) - 2} + \cdots \\ &\quad + \frac{\text{OPT}}{\Omega\left(G\right) - (\Omega\left(G\right) - 2)} \\ &\quad + \frac{\text{OPT}}{\Omega\left(G\right) - (\Omega\left(G\right) - 1)} \\ &\le \left(\ln\left(\Omega\left(G\right)\right) + 1\right) \text{OPT}. \end{aligned} \tag{22}$$

5. Performance Evaluation

5.1. Before the Simulation Setup. Because there is no real data set which is consistent with the proposed system model, we have to mine the ways of human mobility from Gowalla [52] and Brightkite [53], which come from the location-based social networking website where users share their locations by checking-in. The details of the data sets are listed in Table 7.

We consider the variation law of user's mobile preferences because the sensing task is dependent on location in most crowd sensing systems. Observing a user's visiting history can help discover the user's abilities to fulfill the subtasks in different locations.

We divide the region Q into $k * k$ square blocks and let $q(x, y) \in Q$ denote the block in which $x, y \in \{1, 2 \ldots, k\}$ represents horizontal and vertical locations, respectively. Let $f_i(t, q(x, y))$ denote the number of checking-ins of user i during the time period $t \in [t_1, t_2]$ in block $q(x, y)$. If $f_i(t, q(x, y))$ is greater than zero, $q(x, y)$ is called the reachable region of user i during the time period t.

The reachable regions of user i can be viewed as the subtasks in different locations that the user can fulfill. Let $h(i, t)$ denote the number of reachable regions which can be

Table 7: Facts about studied traces.

Trace source	Brightkite	Gowalla
Time/duration of trace	2008/4–2010/10	2009/2–2010/10
The number of users	58228	196591
The number of check-ins	4491143	6442890

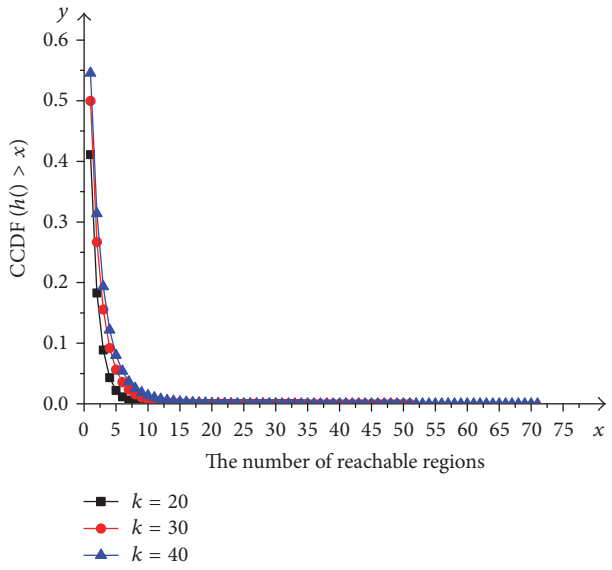


Figure 3: The complementary cumulative distribution of $h(i,t)$ in Gowalla.

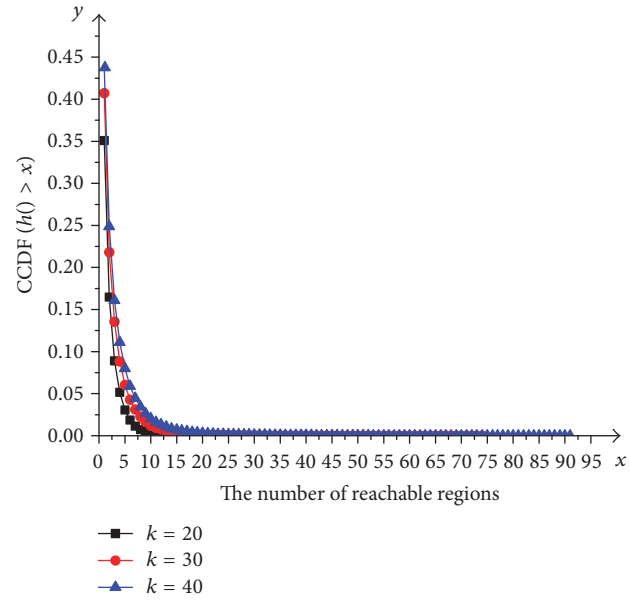


Figure 4: The complementary cumulative distribution of $h(i,t)$ in Brightkite.

viewed as the number of subtasks that the user can fulfill, which is calculated by

$$h(i,t) = \sum_{x=1}^{k} \sum_{y=1}^{k} \eta_i (t, q(x,y)),$$
$$\eta_i (t, q(x,y)) = \begin{cases} 0, & f_i (t, q(x,y)) = 0, \\ 1, & f_i (t, q(x,y)) > 0. \end{cases} \tag{23}$$

With different partition granularity in different data set, more than 95% of the users' number of reachable regions is smaller than 1.25% of the number of region blocks; the complementary cumulative distribution of $h(i,t)$ follows the same power-law distribution formulated by (24), which can be seen in Figures 3 and 4, respectively.

$$p(X > x) = \left(\frac{x}{x_{\min}} \right)^{-k}. \tag{24}$$

The phenomenon of power-law distribution is consistent with our life experience: the activity scope of most people has a limited range in several specific regions in their daily life. The above-mentioned result tells us how to set the number of subtasks that the user can fulfill.

5.2. Performance Evaluation. In order to evaluate QAIM, we first introduce two baseline algorithms with the similar ideas of using redundancy.

(i) max(K) is derived from the K-depth coverage objective solution proposed in [43]. No matter how different the sensing quality factor of each subtask is, K is set to the maximal value of these quality factors.

(ii) Greedy(1) is derived from the idea proposed in [41]. No matter how many subtasks one participant can do, he is only assigned with one subtask at a time. So, the participant is selected in the nondecreasing order of the bid price.

The performance metrics include the social cost, the number of winners, the running time, and the truthfulness, and the importance of reputation score is also checked up.

Simulation parameters are shown in Table 8. Each measurement is averaged over 100 instances.

(1) Impact of $|U|$. Figures 5–7 show the performance of QAIM(S) with different candidates when the number of sensing subtasks is set to 100. As shown in Figures 5 and 6, the social cost and the number of winners of QAIM(S) are both less than those of Greedy(1) or max(K). The variation does not follow the rule of decreasing with increment of the

Table 8: Simulation settings.

Simulation parameters	Settings
$h(g_k)$	Uniformly distributed over [3, 7]
b_{v_i}	Uniformly distributed over [5, 7]
$\psi_{v_i} = \{g_1^{v_i}, g_2^{v_i}, \ldots, g_k^{v_i}\}$	$g_k^{v_i}$ is random in G but $\lvert\psi_{v_i}\rvert$ is between 3 and 10 which abides to Pareto distribution with $x_{\min} = 1$ and $k = 1.75$.

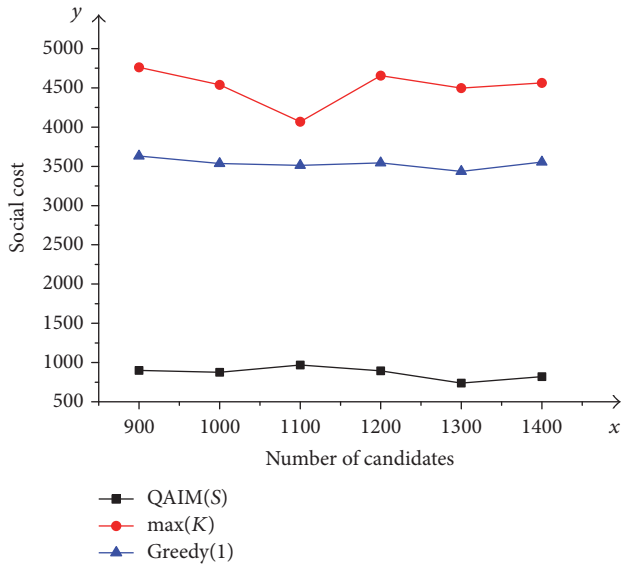


Figure 5: The social cost with different candidates.

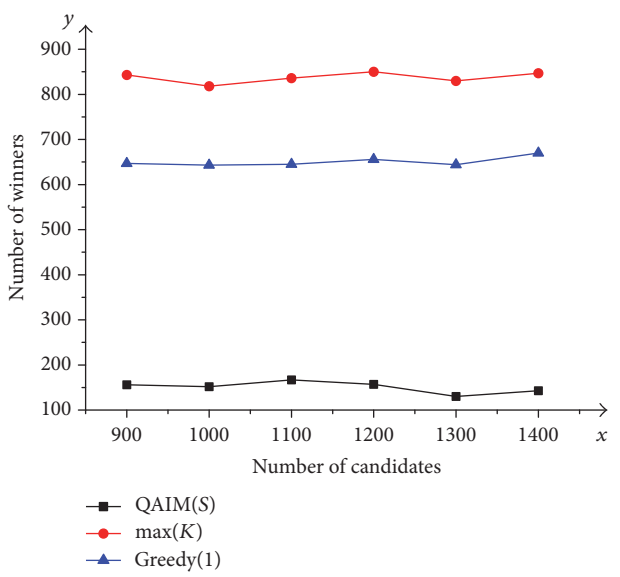


Figure 6: The number of winners with different candidates.

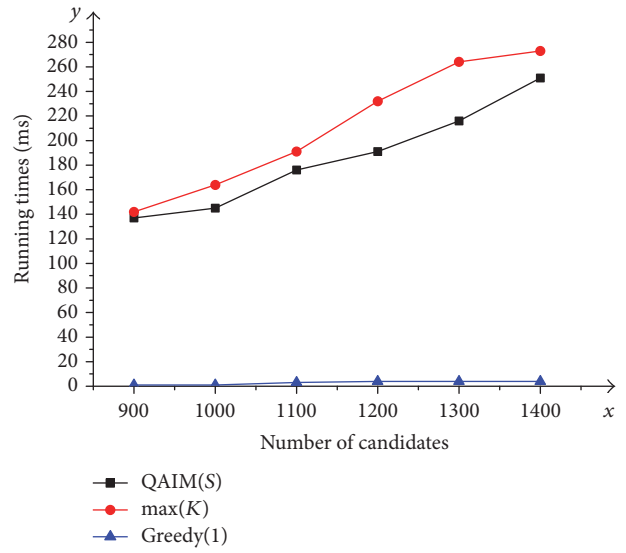


Figure 7: The running time with different candidates.

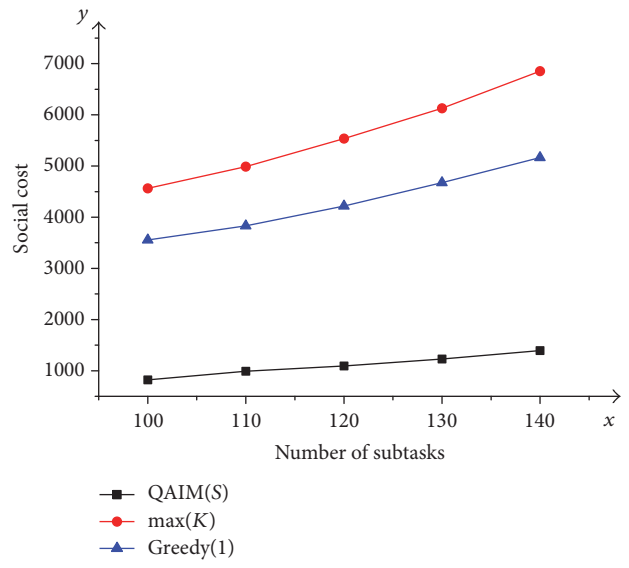


Figure 8: The social cost with different number of sensing subtasks.

number of candidates but is within a certain range. The reason is $h(g_k)$ and b_{v_i} which are generated randomly. QAIM(S) has superiority in achieving high quality crowd sensing with minimum social cost. The running time of QAIM(S) is larger than Greedy(1) but less than max(K) as shown in Figure 7. The variation trend of the running time is consistent with the property of theoretical analysis which increases with the increasing number of candidates.

(2) Impact of $|G|$. Figures 8–10 show the performance with a fixed number of 1400 candidates when the number of sensing subtasks varies from 100 to 140 with increment of 10. As shown in Figures 8 and 9, both the social cost and the number

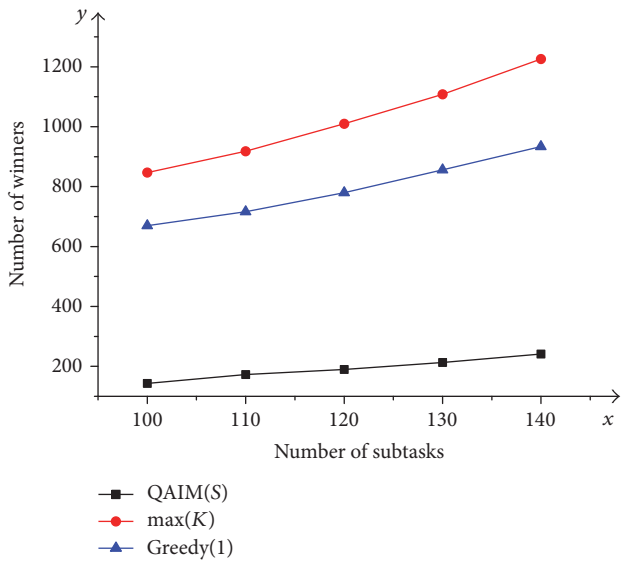


Figure 9: The number of winners with different number of sensing subtasks.

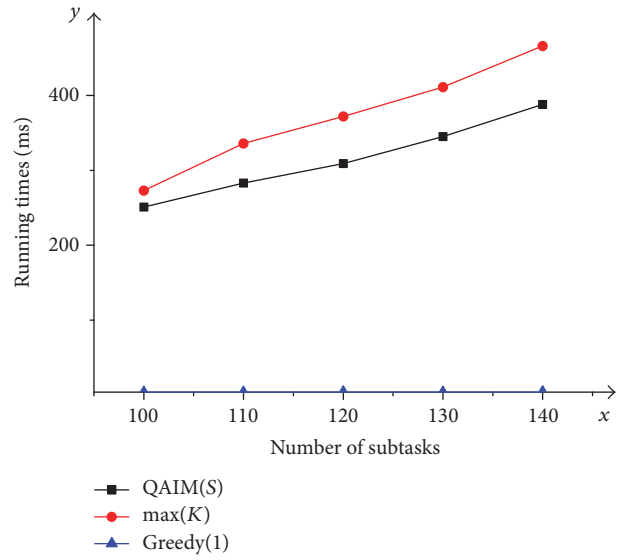


Figure 10: The running time with different number of sensing subtasks.

of winners of QAIM(S) increase with the increment of $|G|$ and are less than other algorithms. The running time of QAIM(S) is larger than Greedy(1) but less than max(K) and increases with $|G|$ as shown in Figure 10, which is consistent with the property of theoretical analysis.

(3) Truthfulness. We verified the truthfulness of QAIM with different candidates when the number of subtasks is set to 100. We randomly selected the 78th participant and changed the bid price b_{78} of the 78th participant. When $b_{78} > p_{78}$, the 78th participant would not be selected. The running time of QAIM(P) is recorded in Figure 11 which shows the time cost of the truthfulness. The running time of QAIM(P) is bounded by 80 and increases with the increment of the number of

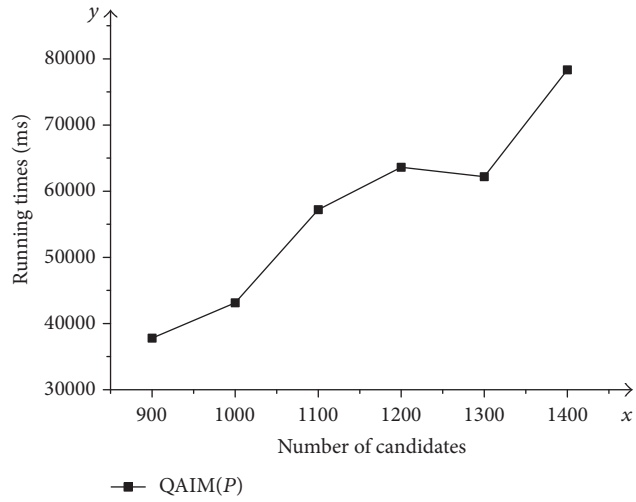


Figure 11: The running time of QAIM(P) with different candidates.

candidates except when the number of candidates is 1300, which is a reasonable phenomenon since the running time of QAIM(P) is related to not only the number of candidates but also the number of winners.

(4) The Effect of Reputation Value. Finally, we verified the importance of the calculation of reputation value. We first set the 78th participant as the malicious user and offer the contrary sensing result to correct ones of all subtasks intentionally; we find that it would not be selected after the second test. Then we reset the reputation score to 1 and let the 78th participant be selected but the 78th participant does not fulfill one of the subtasks; we find that it would be selected after the second test and would not be selected after the third test.

6. Conclusion

In this paper, we address the fundamental research issue: how can we achieve high quality crowd sensing with the minimum social cost? To answer this question, we study different conditions of recruiter and candidates in crowd sensing system. Based on the findings, we formulate the sensing quality assurance problem as an optimization problem (MQMUS) and prove it to be NP-hard. We design a polynomial-time greedy approximation algorithm QAIM which consists of two phases: QAIM(S) selects appropriate participants to satisfy the objective of this research which approximates the optimal solution with the times of $\ln(\Omega(G)) + 1$ and QAIM(P) eliminates the fear of market manipulation. Through rigorous theoretical analysis, we demonstrate the proposed mechanisms with the properties of high computation efficiency, individual rationality, and truthfulness and then evaluate our algorithm using synthetic data with the features of real data sets. Evaluations show that our algorithms outperform existing approaches. In the future work, we will explore the quality-aware incentive mechanisms in more complex scenarios, for example, how to prevent cocheating using the

history of mobility traces and the completed tasks list of participants.

Appendix

The MQMUS Problem Is NP-Hard

Demonstration. In order to prove that the MQMUS problem is NP-hard, we first prove that the MQMUS_1 problem is NP-hard. We define MQMUS_1 as a special case of MQMUS in which every $h(g_k)$ is equal to one. Thereafter, we conclude that the MQMUS problem is NP-hard.

The problem of MQMUS_1 can be illustrated below which is a set cover problem with weight b_i.

Given a set of elements $G = \{g_1, g_2, \ldots, g_k, \ldots, g_e\}$ and a set of $B = \{B_1, B_2, \ldots, B_k, \ldots, B_n\}$ in which ψ_i is the subset of G and b_i is the cost of ψ_i, the problem of MQMUS_1 is to find a collection $S = \{B_{s_1}, B_{s_2}, \ldots, B_{s_k}\}$ from B such that the union of ψ_{s_i} equals G with the least costs. We cannot find an efficient optimal solution for the special case of MQMUS_1 in polynomial time, so MQMUS_1 is NP-hard.

MQMUS_1 is a special instance of MQMUS while $h(g_k)$ varies with different sensing quality requirement. Therefore, MQMUS is also NP-hard.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was sponsored by the National Natural Science Foundation of China (nos. 61572261, 61472193, and 61373138), the Natural Science Foundation of Jiangsu Province (no. BK20141429), Postdoctoral Foundation (no. 2016T90485), the Sixth Talent Peaks Project of Jiangsu Province (nos. DZXX-017 and JNHB-062), and Open Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks (WSNLBZY201518).

References

[1] R. K. Ganti, F. Ye, and H. Lei, "Mobile crowdsensing: current state and future challenges," *IEEE Communications Magazine*, vol. 49, no. 11, pp. 32–39, 2011.

[2] V. Sivaraman, J. Carrapetta, K. Hu, and B. G. Luxan, "HazeWatch: a participatory sensor system for monitoring air pollution in Sydney," in *Proceedings of the IEEE 38th Conference on Local Computer Networks Workshops (LCN '13)*, pp. 56–64, IEEE Computer Society, Sydney, Australia, October 2013.

[3] R. K. Rana, C. T. Chou, S. S. Kanhere, N. Bulusu, and W. Hu, "Ear-phone: an end-to-end participatory urban noise mapping system," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN '10)*, pp. 105–116, Stockholm, Sweden, April 2010.

[4] N. Maisonneuve, M. Stevens, M. E. Niessen, and L. Steels, "NoiseTube: Measuring and mapping noise pollution with mobile phones," in *Information Technologies in Environmental Engineering*, Environmental Science and Engineering, pp. 215–228, Springer Berlin Heidelberg, Heidelberg, Gremany, 2009.

[5] P. Mohan, V. Padmanabhan N, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," Tech. Rep. 323-336, Microsoft Technical Report, Raleigh, NC, USA, 2008.

[6] E. Koukoumidis, L.-S. Peh, and M. R. Martonosi, "SignalGuru: leveraging mobile phones for collaborative traffic signal schedule advisory," in *Proceedings of the 9th International Conference on Mobile Systems, Applications, and Services*, pp. 127–140, ACM, Bethesda, Md, USA, July 2011.

[7] A. Thiagarajan, L. Ravindranath, K. LaCurts et al., "VTrack: accurate, energy-aware road traffic delay estimation using mobile phones," in *Proceedings of the 7th ACM Conference on Embedded Networked Sensor Systems (SenSys '09)*, pp. 85–98, November 2009.

[8] S. Reddy, A. Parker, J. Hyman, J. Burke, D. Estrin, and M. Hansen, "Image browsing, processing, and clustering for participatory sensing: Lessons from a DietSense prototype," in *Proceedings of the 4th Workshop on Embedded Networked Sensors, EmNets 2007*, pp. 13–17, irl, June 2007.

[9] E. Miluzzo, N. D. Lane, K. Fodor et al., "Sensing meets mobile social netwo rks: the design, implementation and evaluati on of the cenceme application," in *Proceedings of the 6th ACM Conference on Embedded Networked Sensor Systems (SenSys '08)*, pp. 337–350, ACM, New York, NY, USA, November 2008.

[10] W. Dong, B. Lepri, and S. Pentland, "Tracking co-evolution of behavior and relationships with mobile phones," *Tsinghua Science and Technology*, vol. 17, no. 2, pp. 136–151, 2012.

[11] Z. Yang, C. Wu, and Y. Liu, "Locating in fingerprint space: wireless indoor localization with little human intervention," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (Mobicom '12)*, pp. 269–280, ACM, Istanbul, Turkey, August 2012.

[12] C. Costa, C. Laoudias, D. Zeinalipour-Yazti, and D. Gunopulos, "SmartTrace: Finding similar trajectories in smartphone networks without disclosing the traces," in *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering, ICDE 2011*, pp. 1288–1291, April 2011.

[13] S. Matyas, C. Matyas, C. Schlieder, P. Kiefer, H. Mitarai, and M. Kamata, "Designing location-based mobile games with a purpose - Collecting geospatial data with cityexplorer," in *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology, ACE 2008*, pp. 244–247, December 2008.

[14] Sensorly. http://www.sensorly.com.

[15] S. B. Eisenman, E. Miluzzo, N. D. Lane, R. A. Peterson, G.-S. Ahn, and A. T. Campbell, "The BikeNet mobile sensing system for cyclist experience mapping," in *Proceedings of the 5th International Conference on Embedded Networked Sensor Systems (SenSys '07)*, pp. 87–101, Sydney, Australia, November 2007.

[16] Y. Liu, Y. Zhao, L. Chen, J. Pei, and J. Han, "Mining frequent trajectory patterns for activity monitoring using radio frequency tag arrays," *IEEE Transactions on Parallel and Distributed Systems*, vol. 23, no. 11, pp. 2138–2149, 2012.

[17] P. Zhou, Y. Zheng, and M. Li, "How long to wait?: predicting bus arrival time with mobile phone based participatory sensing," in *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services (MobiSys '12)*, pp. 379–392, ACM, June 2012.

[18] L. Barkhuus, M. Chalmers, P. Tennent et al., "Picking Pockets on the Lawn: The Development of Tactics and Strategies in a Mobile Game [C]," in *Proceedings of UBICOMP*, pp. 358–374, 2005.

[19] C. Schlieder, "Representing the meaning of spatial behavior by spatially grounded intentional systems," *Lecture Notes in Computer Science*, vol. 3799, pp. 30–44, 2005.

[20] M. Bell, M. Chalmers, L. Barkhuus et al., "Interweaving mobile games with everyday life," in *Proceedings of the CHI 2006: Conference on Human Factors in Computing Systems*, pp. 417–426, April 2006.

[21] P. Kiefer, S. Matyas, and C. Schlieder, "Playing on a line: Location-based games for linear trips," in *Proceedings of the 4th International Conference on Advances in Computer Entertainment Technology, ACE 2007*, pp. 250-251, June 2007.

[22] N. Bulusu, C. Chou T, S. Kanhere et al., "Participatory Sensing in Commerce: Using Mobile Phones to Track Market Price Dispersion," *Proc. Urbansense*, pp. 6–10, 2008.

[23] L. Deng and L. P. Cox, "Live compare: grocery bargain hunting through participatory sensing," in *Proceedings of the 10th Workshop on Mobile Computing Systems and Applications (HotMobile '09)*, pp. 1–6, ACM, New York, NY, USA, February 2009.

[24] B. L. Sullivan, C. L. Wood, M. J. Iliff, R. E. Bonney, D. Fink, and S. Kelling, "eBird: a citizen-based bird observation network in the biological sciences," *Biological Conservation*, vol. 142, no. 10, pp. 2282–2292, 2009.

[25] M. Bell, S. Reeves, B. Brown et al., "EyeSpy," in *Proceedings of the the SIGCHI Conference*, p. 123, Boston, MA, USA, April 2009.

[26] E. Kanjo, "NoiseSPY: a real-time mobile phone platform for urban noise monitoring and mapping," *Mobile Networks and Applications*, vol. 15, no. 4, pp. 562–574, 2010.

[27] K. Han, E. A. Graham, D. Vassallo, and D. Estrin, "Enhancing motivation in a mobile participatory sensing project through gaming," in *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust, PASSAT 2011 and 2011 IEEE International Conference on Social Computing, SocialCom 2011*, pp. 1443–1448, October 2011.

[28] B. Hoh, T. Yan, D. Ganesan, K. Tracton, T. Iwuchukwu, and J.-S. Lee, "TruCentive: a game-theoretic incentive platform for trustworthy mobile crowdsourcing parking services," in *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC '12)*, pp. 160–166, IEEE, Anchorage, Alaska, USA, September 2012.

[29] K. O. Jordan, I. Sheptykin, B. Grüter, and H.-R. Vatterrott, "Identification of structural landmarks in a park using movement data collected in a location-based game," in *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Computational Models of Place, COMP 2013*, pp. 1–8, November 2013.

[30] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM '12)*, pp. 1701–1709, Orlando, Fla, USA, March 2012.

[31] D. Yang, G. Xue, X. Fang et al., "Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking (MobiCom '12)*, pp. 173–184, ACM, August 2012.

[32] J.-S. Lee and B. Hoh, "Sell your experiences: A market mechanism based incentive for participatory sensing," in *Proceedings of the 8th IEEE International Conference on Pervasive Computing and Communications, PerCom 2010*, pp. 60–68, April 2010.

[33] Z. Feng, Y. Zhu, Q. Zhang, L. M. Ni, and A. V. Vasilakos, "TRAC: Truthful auction for location-aware collaborative sensing in mobile crowdsourcing," in *Proceedings of the 33rd IEEE Conference on Computer Communications, IEEE INFOCOM 2014*, pp. 1231–1239, May 2014.

[34] X. Zhang, G. Xue, R. Yu, D. Yang, and J. Tang, "Truthful incentive mechanisms for crowdsourcing," in *Proceedings of the 34th IEEE Annual Conference on Computer Communications and Networks, IEEE INFOCOM 2015*, pp. 2830–2838, May 2015.

[35] J. Xu, J. Xiang, and D. Yang, "Incentive Mechanisms for Time Window Dependent Tasks in Mobile Crowdsensing," *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6353–6364, 2015.

[36] J. Xu, J. Xiang, and Y. Li, "Incentivize maximum continuous time interval coverage under budget constraint in mobile crowd sensing," *Wireless Networks*, vol. 23, no. 5, pp. 1–14, 2017.

[37] A. Subramanian, G. Kanth S, and R. Vaze, "Offline and Online Incentive Mechanism Design for Smart-phone Crowdsourcing," *Computer Science*, pp. 1–12, 2013.

[38] D. Zhao, X.-Y. Li, and H. Ma, "How to crowdsource tasks truthfully without sacrificing utility: Online incentive mechanisms with budget constraint," in *Proceedings of the 33rd IEEE Conference on Computer Communications, IEEE INFOCOM 2014*, pp. 1213–1221, May 2014.

[39] L. Jaimes G, I. Vergara-Laurens, and M. Labrador, "A Location-Based Incentive Mechanism for Participatory Sensing Systems with Budget Constraints," *Dissertations Theses-Gradworks*, vol. 25, no. 3, pp. 103–108, 2012.

[40] H. Jin, L. Su, D. Chen, K. Nahrstedt, and J. Xu, "Quality of information aware incentive mechanisms for mobile crowd sensing systems," in *Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing (MobiHoc '15)*, pp. 167–176, ACM, Hangzhou, China, June 2015.

[41] X. Zhang, Z. Yang, Y. Gong, Y. Liu, and S. Tang, "Spatial Recruiter: Maximizing sensing coverage in selecting workers for spatial crowdsourcing," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, 2017.

[42] J. Xu, H. Li, Y. Li, D. Yang, and T. Li, "Incentivizing the Biased Requesters: Truthful Task Assignment Mechanisms in Crowdsourcing," in *Proceedings of the 2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9, San Diego, Calif, USA, June 2017.

[43] H. Xiong, D. Zhang, G. Chen, L. Wang, V. Gauthier, and L. E. Barnes, "ICrowd: Near-Optimal Task Allocation for Piggyback Crowdsensing," *IEEE Transactions on Mobile Computing*, vol. 15, no. 8, pp. 2010–2022, 2016.

[44] Y. Wang, X. Jia, Q. Jin, and J. Ma, "QuaCentive: a quality-aware incentive mechanism in mobile crowdsourced sensing (MCS)," *Journal of Supercomputing*, vol. 72, no. 8, pp. 2924–2941, 2016.

[45] Z. He, J. Cao, and X. Liu, "High quality participant recruitment in vehicle-based crowdsourcing using predictable mobility," in *Proceedings of the 34th IEEE Annual Conference on Computer Communications and Networks, IEEE INFOCOM 2015*, pp. 2542–2550, May 2015.

[46] D. Fudenberg and J. Tirole, *Game Theory*, MIT Press, 1991.

[47] V. Krishna, *Auction Theory*, Academic Press, 2009.

[48] C. Tanas and J. Herrera-Joancomartí, "When users become sensors: Can we trust their readings?" *International Journal of Communication Systems*, vol. 28, no. 4, pp. 601–614, 2015.

[49] L. Kazemi, C. Shahabi, and L. Chen, "GeoTruCrowd: Trustworthy query answering with spatial crowdsourcing," in *Proceedings*

of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM SIGSPATIAL GIS 2013, pp. 304–313, ACM, November 2013.

[50] James., *The Wisdom of Crowds: Why The Many Are Smarter than The Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*, Doubleday, 2004.

[51] R. B. Myerson, "Optimal auction design," *Mathematics of Operations Research*, vol. 6, no. 1, pp. 58–73, 1981.

[52] Gowalla [EB/OL]. http://snap.stanford.edu/data/loc-gowalla.html

[53] Brightkite [EB/OL]. http://snap.stanford.edu/data/loc-brightkite.html.

Hindawi
Journal of Sensors
Volume 2017, Article ID 7360953, 18 pages
https://doi.org/10.1155/2017/7360953

Research Article

Generating Human-Like Velocity-Adapted Jumping Gait from sEMG Signals for Bionic Leg's Control

Weiwei Yu,[1,2] Weihua Ma,[3] Yangyang Feng,[1] Runxiao Wang,[1] Kurosh Madani,[4] and Christophe Sabourin[4]

[1] *School of Mechanical Engineering, Northwestern Polytechnical University, Xi'an 710072, China*
[2] *Unmanned System Research Institute, Northwestern Polytechnical University, Xi'an 710072, China*
[3] *National Key Laboratory of Aerospace Flight Dynamics, School of Astronautics, Northwestern Polytechnical University, Xi'an 710072, China*
[4] *Images, Signals and Intelligence Systems Laboratory (LISSI/EA 3956), Senart-Fontainebleau Institute of Technology, UPEC, Bât. A, 77127 Lieusaint, France*

Correspondence should be addressed to Weiwei Yu; yuweiwei@nwpu.edu.cn

Received 11 March 2017; Revised 5 June 2017; Accepted 4 July 2017; Published 29 August 2017

Academic Editor: Xiaokun Zhang

In the case of dynamic motion such as jumping, an important fact in sEMG (surface Electromyogram) signal based control on exoskeletons, myoelectric prostheses, and rehabilitation gait is that multichannel sEMG signals contain mass data and vary greatly with time, which makes it difficult to generate compliant gait. Inspired by the fact that muscle synergies leading to dimensionality reduction may simplify motor control and learning, this paper proposes a new approach to generate flexible gait based on muscle synergies extracted from sEMG signal. Two questions were discussed and solved, the first one concerning whether the same set of muscle synergies can explain the different phases of hopping movement with various velocities. The second one is about how to generate self-adapted gait with muscle synergies while alleviating model sensitivity to sEMG transient changes. From the experimental results, the proposed method shows good performance both in accuracy and in robustness for producing velocity-adapted vertical jumping gait. The method discussed in this paper provides a valuable reference for the sEMG-based control of bionic robot leg to generate human-like dynamic gait.

1. Introduction

Surface Electromyogram (sEMG) is the electrical manifestation of muscular contractions, which reflexes plentiful neural control information. For its strong relationship with human motion pattern, sEMG has been taken as an ideal noninvasive control signal for exoskeletons [1], myoelectric prostheses [2], and biorobot [3] and moreover for the development of rehabilitation robots [4]. An important factor presenting in the sEMG-based biomechanical leg control, during the dynamic motion such as jumping and running, is the fact that multichannel sEMG signals contain mass data and vary greatly with time. Such makes it more difficult to generate compliant motion. It is more desired to generate compliance gait with sEMG signal, for exoskeletons, myoelectric prostheses, and rehabilitation gait.

It is hypothesized that the CNS (central nervous system) coordinates groups of muscles with specific activation balances and temporal profiles, to simplify the generation of intricate movements [5]. These building modules, known as muscle synergies, can be used as a small number of coactivation patterns to imitate the performance of movement [6]. It is very attractive to point out that these synergies make it possible for the motor intentions to be rapidly translated into muscle activation and the systems can learn and plan movements so fast [7, 8]. From the computational perspective, with muscle synergies leading to dimensionality reduction that simplifies motor control and learning, such observation has recently raised the interest of many researchers to develop control strategies in robotic and biomechanical application [9]. So, we suppose that it is reasonable to apply the muscle synergies to simplify the generation of compliant dynamic

gait, such as hopping, based on sEMG for biomechanical leg control.

The generality of muscle synergies across different motor tasks has been illustrated in human walking, running [10], and forward and backward pedaling movement [11]. For example, previous work on human balance control has shown that the same muscle synergies can account for balance responses under different dynamic conditions: stepping and nonstepping postural responses [12]. Five modules satisfy the human walking in sagittal plane, while sixth module, which contributes primarily to mediolateral balance control and contralateral leg swing, is needed to satisfy the additional nonsagittal plane demands of 3D walking [13]. Although the same muscle synergies are used across multiple tasks, in some instances, new synergies may be recruited to accomplish a new behavior goal [14]. Like in the motion of human finger spelling, the recruitment of muscle synergies is correlated with common hand postures [15]. Thus, the first problem that would be discussed in this paper is whether the same set of muscle synergies can explain the different phases of dynamic gait with various velocities.

To generate the gait pattern with muscle synergies is related to the problem of identifying the forward relationship between sEMG and resultant joint movement. In the past few years, several contributions were proposed to predict the forward relationship between synergies and joint angles, joint torques, or end force. Like the approaches based on the linear models, Artemiadis and Kyriakopoulos used Linear Time-Invariant (LTI) model to take PCA synergy features as inputs to relate synergy features to anthropomorphic joint movements in three-dimensional space [16]. In order to predict the wrist intended activation of natural movement, synergies of both DOF (Degree of Freedom) were extracted at once, and three synergies can achieve simultaneous control when applied separately on each DOF [17].

Nonlinear model based approaches are not as reliant on robust synergy features as linear models and therefore are able to represent rational complex relationship between synergies and desired outputs. Hahne et al. systematically compared linear and nonlinear regression techniques for an independent, simultaneous, and proportional myoelectric control of wrist movements and got the results that the kernel ridge regression outperformed the other methods, but with higher computational costs [18]. A hybrid time-delayed artificial neural network was investigated to predict clenching movements during mastication from sEMG signals. Actual jaw motions and sEMG signals from the masticatory muscles were recorded and used as output and input, respectively, [19]. Muceli and Farina also adopted multiplayer perceptron (MLP) artificial neural networks to estimate kinematics of the hand wrist from EMG signals of the contralateral limb during mirrored bilateral movements in free space [20].

However, during the dynamic motion such as jumping and running, the multichannel sEMG signals vary greatly with time which may result in high risk for overfitting models to training data and frequent retraining [21]. Moreover, for the application of exoskeletons, myoelectric prostheses, and rehabilitation gait, it is always desirable to generate self-adapted gait with limited experimental data. Therefore, the second problem that needed to be solved in this paper is how to generate flexible continuous dynamic gait with limited sets of experimental data based on these extracted muscle synergies, while avoiding overfitting model and alleviating model sensitivity to sEMG transient changes.

Since vertical jumping is the fundamental movement pattern of dynamic motion, such as jumping, bouncing, and running, based on the above discussion, we will focus on generating flexible gait of vertical jumping with sEMG signals for biomechanical leg. The paper is organized as follows: In Section 2, experimental protocol and vertical jumping motion are introduced. How to extract muscle synergies is explained in Section 3. To generate the coordinating variation of joint angles, a synergy-based computational framework built on the Fuzzy Wavelet Neural Networks is introduced in Section 4. Section 5 shows the experimental simulation results of the generalization gaits for vertical jumping motion. The advantages of this approach are the possibilities to generate a self-adaptive gait pattern of vertical jumping with limited number of experimental data, which is very meaningful for the sEMG-based robotic leg application.

2. Experimental Protocol for Vertical Jumping

Six male volunteers with no lower-limb functional limitations or neuromuscular disorders participated in the study. The participants' physical characteristics are the following: age, 23 ± 1.5 years; height, 1.77 ± 0.03 m; and body mass, 72.5 ± 10.0 kg (mean $\pm$ SD). The protocol was approved by the local ethical committee and accorded with the guidelines set out in the Declaration of Helsinki (1983).

2.1. Experimental Protocol. Six volunteer subjects were trained for vertical jumping with uniform and variable speed before the experiment. The experiment contained 16 trials. For the first 8 trials, each containing 15 jumps, the subjects were asked to hop with fast speed for 4 trials and continuously jump with slow speed in the next 4 trials. In the following 8 trials, each containing 20 jumps, for the first 4 trials the subjects were asked to continuously hop with five fast and five slow hops, which were carried out alternatively. For the last 4 trials, slow speed and fast speed hops were carried alternatively in the same way. In this case, not only the muscles coactivities of jumping motion with different velocities but also the muscle activation for changing velocity could be recorded.

Three-dimensional kinematic data were collected using VICON 10-camera motion capture system at a frequency of 500 Hz. AMTI 3D force platform was used to keep track of plantar force with sampling frequency of 500 Hz. The BioVision 8 channel sEMG device with sampling frequency of 1000 Hz was used to record the activities of seven leg muscles: tibialis anterior, gastrocnemius, soleus, vastus medialis, rectus femoris, gluteus maximus, and biceps femoris. The eighth channel of sEMG equipment was used for synchronization with motion capture data. The experimental environment is shown in Figure 1.

Figure 1: Experimental settings and environment.

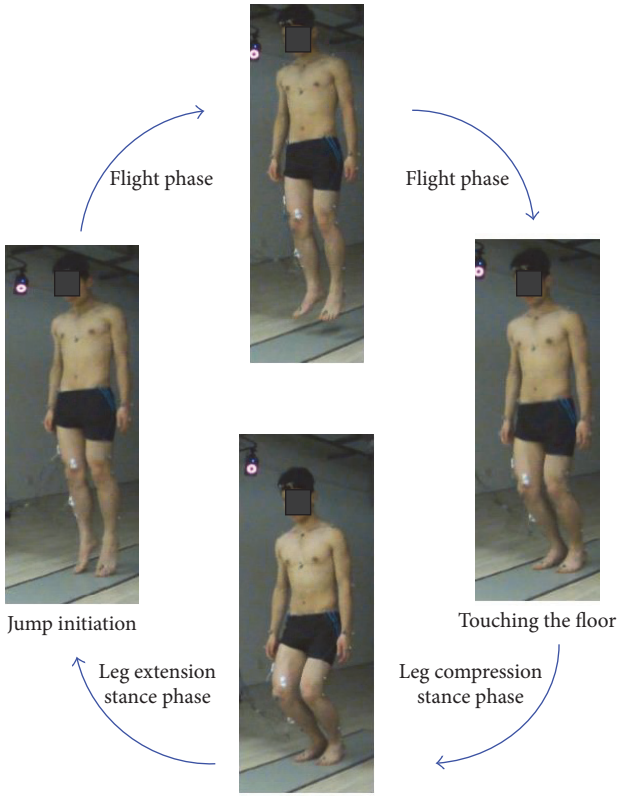
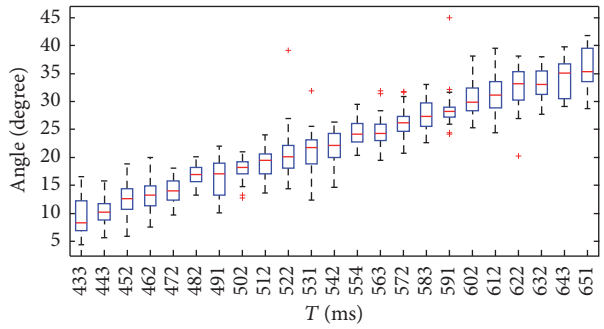


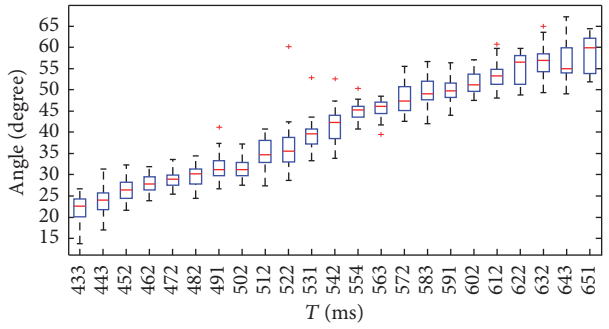
Figure 2: Phase transition of vertical jumping movement.

2.2. Vertical Jumping Movement

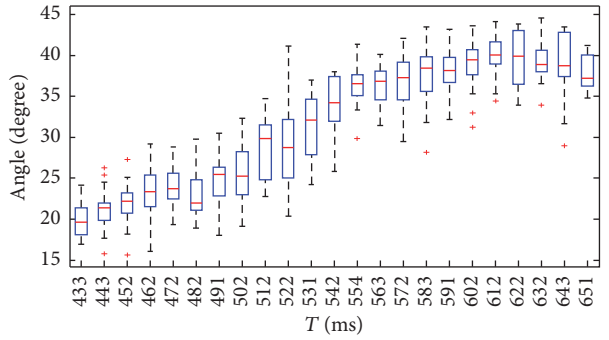
2.2.1. Phase Analysis of Vertical Jumping Movement. Because of the changing of impact force imposed on the pelma, the jumping movement is usually divided into stance phase and flight (swing) phase. Figure 2 describes the transition of each phase. During the stance phase, all the joint angles firstly compress. This compression stage begins at the moment of foot touching the floor and lasts until the vertical velocity equals zero. This process ensures the leg absorbing the ground impact under high speed and being ready for the next hopping movement. Then, the leg extends to provide enough force for the requirement of flight phase. This extension stage ends until there is no ground reaction force. Based on the experiment results, it is important to point out that the maximum plantar force always appears after the moment



(a) Maximum hip compression angles



(b) Maximum knee compression angles



(c) Maximum ankle compression angles

Figure 3: Maximum joint compression angles change along with jumping rhythm.

of maximum joint compression. During the flight (swing) phase, the joint angles change to maintain the whole body in balance state and jump with certain velocity.

2.2.2. Influence of Jumping Velocity on Vertical Jumping Gait. To evaluate if the velocity exerts important influence on jumping gait, we will discuss the relevance of jumping rhythm with the joint angles. Taking the motion capture data of one volunteer as an example, the available vertical jumping rhythm is between 430 ms and 660 ms after all the experimental trials finished. The joint angles were grouped according to the same rhythm for each 10 ms interval, and all the experimental data could be divided into 23 groups.

The box-whisker plot is chosen to estimate the maximum joint compression angles along with the changing jumping rhythm. In Figure 3, red line indicates the median of the max compression angle, blue box expresses the interquartile range, and the red + is the extreme outliers. The advantage of using box-whisker plot is that it can minimize the effect of extreme experimental data on the statistical analysis, express data dispersion degree, and be used for comparing examples. From the results it is obvious that the maximum compression angles of all the three joints decrease with jumping cadence increasing, which indicates that the changing velocities have greatly influenced the hopping gait. Therefore, we focus on how to generate the self-adapted jumping gait with variable velocity for biomechanical leg control.

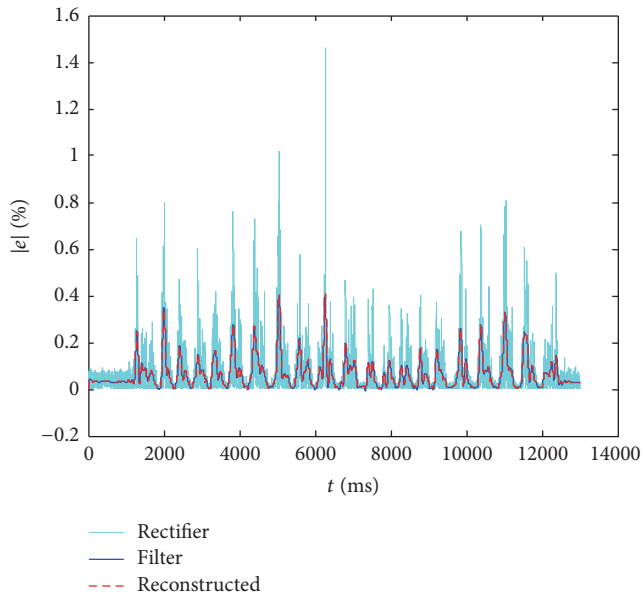


Figure 4: Preprocessing results of one-channel sEMG signals after each digital operation.

3. Muscle Synergies Extraction and Analysis

A prominent hypothesis suggests that the biological system underlies muscle contractions during movement execution in a modular fashion by the CNS. These modular have been observed in forms of muscle synergies. Evidences observed in many cases and species show that regularities of the muscle synergies appear to be very similar across subjects and motor tasks [22, 23]. This section will discuss how to extract muscle synergies and investigate if the same group of synergies can explain the difference phases of vertical jumping motion.

3.1. sEMG Signal Preprocessing and Analysis. Many noise signals will contaminate the original sEMG signals during the experiment, such as the inherent noise of equipment like DC bias, motion artifact, or firing rate of the motor units. To remove these unwanted noises is very important to obtain reliable motion intention of human leg. Specifically in this paper, the sEMG preprocessing experienced the following digital operations:

(i) After applying low-pass filter with 35 Hz to the raw sEMG signals, DC offset was removed for each channel.

(ii) The natural sEMG signal time series vibrates very frequently at the zero point. Through full-wave rectification, the amplitude of signals was more clearly presented.

(iii) In this experiment, the sampling frequency of sEMG signal is 1000 Hz, while the motion capture system samples are at a frequency of 500 Hz; thus, the sEMG signals were subsampled to be consistent with motion signals.

(iv) A high-pass filter was constructed for a cut-off frequency of 10 Hz.

After the above procedures, the noises of the raw sEMG signals are removed and their envelopes are smoothed. The reprocessing results of one channel sEMG after each digital operation are shown in Figure 4. The envelopes of sEMG signals after preprocessing for seven channels are listed in Figure 5, which can be used for joint motion prediction.

3.2. Extracting Muscle Synergies from Different Phases of Jumping

3.2.1. Muscle Synergies Extraction by NMF. Several feature projection techniques, such as Nonnegative Matrix Factorization (NMF) [24], Principal Component Analysis (PCA) [25], Independent Component Analysis (ICA) [26], Linear Discriminant Analysis (LDA) [27], and nonlinear projections [28], can be used to extract muscle coordination pattern. NMF is commonly used as descriptive measure of specific time-invariant muscle synergies because of relaxed constraints on orthogonality and statistical independence between each component and relative robustness to noisy data [24]. NMF is applied to extract synergies of multivariate sEMG data in the following manner. If the structure of generators, which are combined to generate command sEMG signal across tasks, is chosen as the form of spatial dimensionality [29], for K generators,

$$\mathbf{x}^r(t) = \sum_{k=1}^{K} c_k^r(t)\,\mathbf{w}_k, \tag{1}$$

where $\mathbf{x}^r(t)$ are the set of signals ($t = 1 \cdots m$) for task condition r, $c_k^r(t)$ is condition dependent, time-varying combination coefficient for the kth generator, and $\mathbf{w}_k$ is the condition-independent, time-invariant kth spatial generator. Fundamentally, a muscle synergy consists of a time-invariant

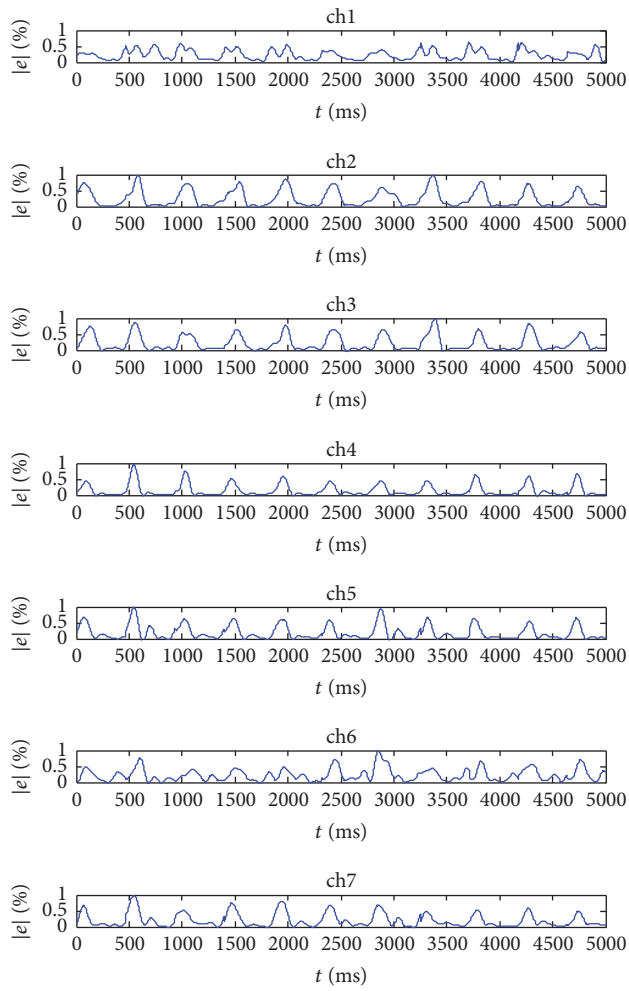


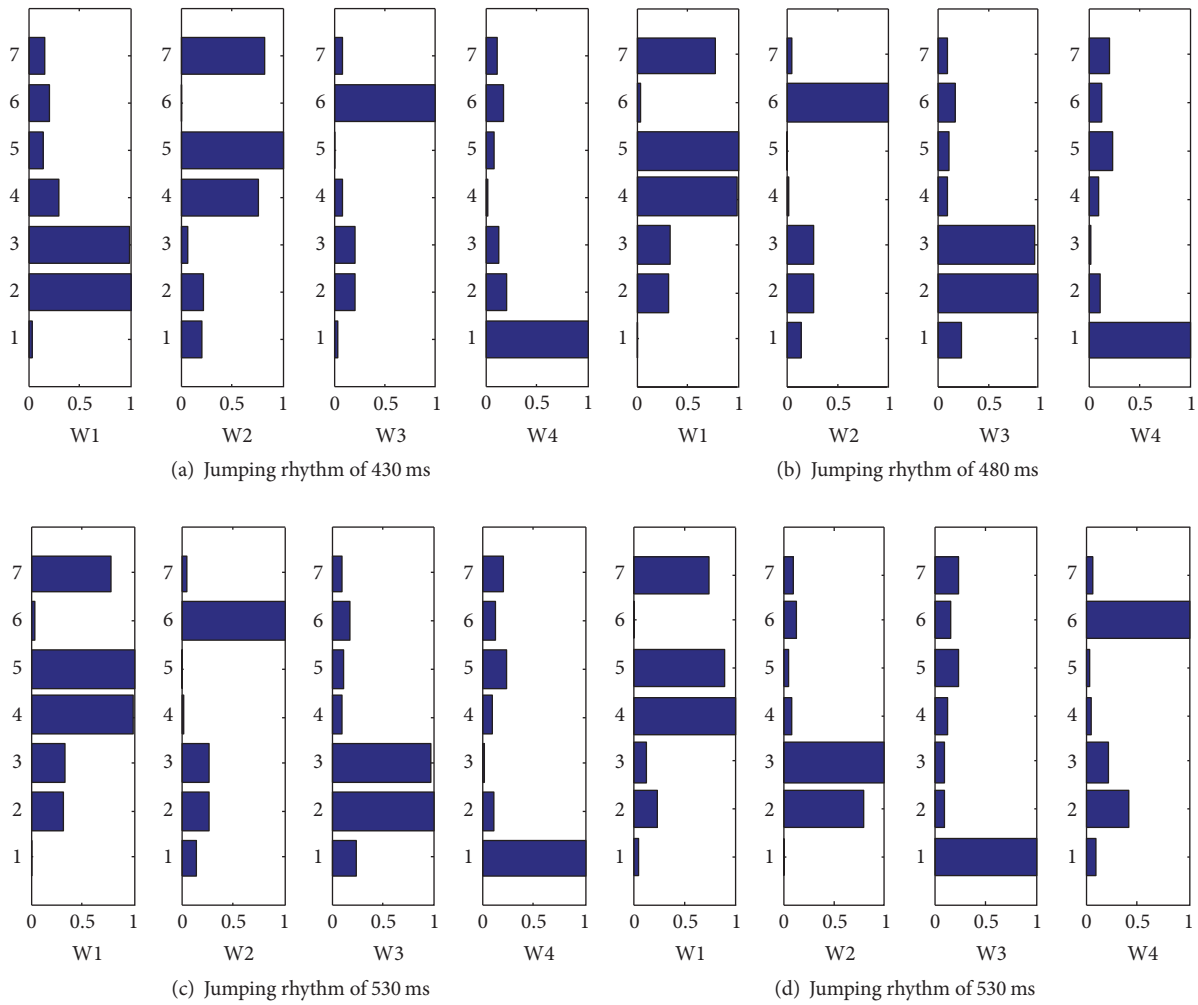
Figure 5: Preprocessing results of sEMG signals for seven channels. ch1, tibialis anterior; ch2, gastrocnemius; ch3, soleus; ch4, vastus medialis; ch5, rectus femoris; ch6, biceps femoris; ch7, gluteus maximus.

weighing coefficient $\mathbf{w}_k$ and a time-varying activation coefficient $c_k^r(t)$. The weighing coefficients within a synergy determine the number of muscles along with the extent of their activation, while the activation coefficient captures when the muscles are active during a task.

3.2.2. Muscle Synergies for Stance and Flight Phase. All the experimental data of each subject were grouped according to the jumping rhythm. The data were divided into five sections with different rhythm $[v1\ \ v2\ \ v3\ \ v4\ \ v5] = [430\text{ ms}\ \ 480\text{ ms}\ \ 530\text{ ms}\ \ 580\text{ ms}\ \ 630\text{ ms}]$. If the number of muscle coordination patterns is chosen as four, the extracted muscle synergies of stance phase for different velocities are given in Figure 6 (taking jumping rhythm of 430 ms, 480 ms, 530 ms, and 580 as examples). The synergies weighting coefficients indicate that the four coactivation patterns are vastus medialis, rectus femoris, and gluteus maximus; gastrocnemius, soleus; tibialis anterior; and biceps femoris, respectively. To evaluate the quality of the extracted synergies, the variance accounted for (VAF) is usually used to calculate the percentage of variability in the sEMG dataset that is accounted for by the extracted synergies. The average VAF value of the six subjects who took part in the experiment is calculated. VAF for leg stance phase with different jumping cadence are listed in Table 1, in which all VAF values are bigger than 90%. This indicates that synergies from one velocity can well explain other velocities and the recorded sEMGs with different jumping velocities are well reconstructed by the same extracted synergies.

Table 1: VAF for stance phase of different jumping rhythm with 4 synergies.

VAF	$v1$	$v2$	$v3$	$v4$	$v5$
$v1$	0.9862	0.9250	0.9151	0.9393	0.9187
$v2$	0.9200	0.9842	0.9554	0.9700	0.9536
$v3$	0.9123	0.9592	0.9906	0.9802	0.9187
$v4$	0.9242	0.9644	0.9769	0.9916	0.9235
$v5$	0.9196	0.9226	0.9006	0.9373	0.9766



(a) Jumping rhythm of 430 ms (b) Jumping rhythm of 480 ms

(c) Jumping rhythm of 530 ms (d) Jumping rhythm of 530 ms

Figure 6: Synergies weighing coefficients of different jumping rhythm for stance phase. 1, tibialis anterior; 2, gastrocnemius; 3, soleus; 4, vastus medialis; 5, rectus femoris; 6, biceps femoris; 7, gluteus maximus.

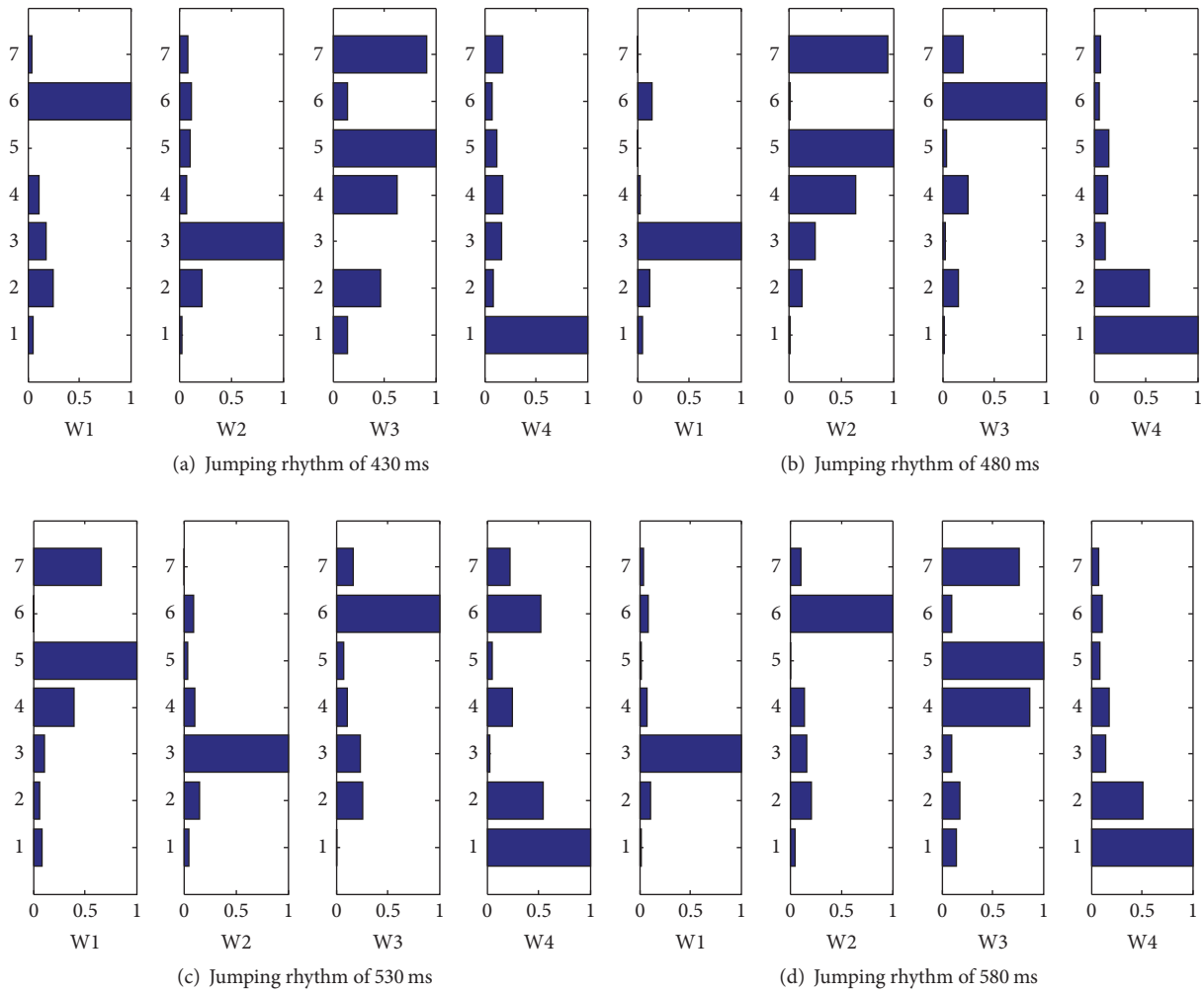
Table 2: VAF for flight (swing) phase of different jumping rhythm with 4 synergies.

VAF	$v1$	$v2$	$v3$	$v4$	$v5$
$v1$	0.9576	0.8837	0.9259	0.8215	0.8697
$v2$	0.9060	0.9651	0.9215	0.8171	0.8387
$v3$	0.9136	0.8835	0.9755	0.8661	0.8837
$v4$	0.8486	0.8324	0.8604	0.9490	0.9002
$v5$	0.8781	0.8830	0.9216	0.9052	0.9402

Next, we will test if the extracted synergies of stance phase can explain the muscle coordination pattern of flight phase. Supposing that the number of synergies is four, the VAF values for different hopping velocities are listed in Table 2. The low VAF values demonstrate that four muscle synergies are not sufficient to explain the flight phase with different velocities and it may employ more muscle coordination patterns. This conclusion is also approved by the calculation results in Figure 7, in which the synergies weighting coefficients are not the same for different velocities. On one hand, during the flight (swing) phase, the number of synergies is more than four and the muscles are more freely organized based on the above results. One the other hand, the effect of changing joint angles is to maintain the whole body in balance state since there is no contact force. Therefore, it is inappropriate to generate the hopping gait of flight phase with the same synergies of the stance phase. In the following section, the muscle synergies are employed to generate hopping gait only for stance phase, while the joint angles trajectory for swing phase is generated by interpolation of the jump-initiation state and touching-the-floor state.

4. Generate Velocity-Adapted Flexible Jumping Gait

4.1. Estimating the Reference Gait Pattern for Stance Phase. The wavelet neural network (WNN) is adopted to identify the relationship between the muscles coactivation patterns



(a) Jumping rhythm of 430 ms (b) Jumping rhythm of 480 ms

(c) Jumping rhythm of 530 ms (d) Jumping rhythm of 580 ms

FIGURE 7: Synergies weighing coefficients of different jumping rhythm for flight phase. 1, tibialis anterior; 2, gastrocnemius; 3, soleus; 4, vastus medialis; 5, rectus femoris; 6, biceps femoris; 7, gluteus maximus.

and coordinated variation of joint angles. The wavelet and the neural network processing can be performed separately. Firstly, the input signals u_i are decomposed using some wavelet basis stored in the hidden layer. The hidden layer consists of neurons, which are usually referred to as wavelons, and their activation functions are drawn from a wavelet basis. Secondly, the wavelet coefficients are output to some summers, whose input weights are updated in accordance with certain learning algorithm.

For a multi-input multioutput nonlinear identification system, the output can be defined by wavelons as

$$y_\vartheta(x) = \sum_{i=1}^{M} w_i \sqrt{\lambda_i}\phi(\lambda_i x - t_i), \tag{2}$$

where λ and t are the dilation and translation parameters, respectively. The parameters $\overline{y}$, w_i, t_i, and λ_i can be grouped into a parameter vector ϑ. The objective function to be minimized is defined as

$$e(\vartheta) = \frac{1}{2} E\left[\left(y_\vartheta(x) - f(x)\right)^2\right]. \tag{3}$$

The minimization of the above function is performed using stochastic gradient algorithm. This recursively modifies ϑ. After each sample pair $\{x_k, f(x_k)\}$, the objective function is written as

$$b(\vartheta, x_k, f(x_k)) = \frac{1}{2}\left(y_\vartheta(x_k) - f(x_k)\right)^2. \tag{4}$$

The parameters t_i and λ_i can be fixed at initialization of the network. w_i, which is the only parameter that needed to be adjusted, is modified in the opposite direction of $e(\vartheta, x_k, f(x_k))$, and the gradient can be computed by the partial derivatives as follows:

$$\frac{\partial e}{\partial w_i} = e_k \sqrt{\lambda_i}\phi(z_{ki}) \tag{5}$$

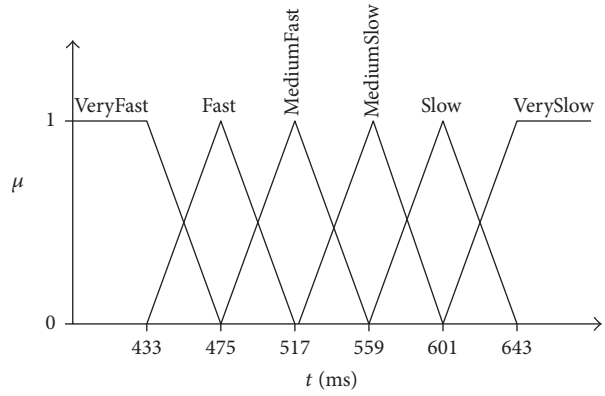


Figure 8: Membership functions for different jumping rhythm.

in which

$$e_k = y_\vartheta(x_k) - f(x_k)$$
$$z_{ki} = \lambda_i x_k - t_i. \tag{6}$$

As it is analyzed in Section 3, the number of muscle synergies which can explain the muscle coactivation patterns of stance phase is four. A 4-input and 3-output wavelet neural network is constructed to predict the coordinating variation of three joint angles with muscle synergies for stance phase, taking the muscle activation coefficient as input signal.

4.2. Estimating the Reference Gait Pattern for Flight Phase. Considering the flight phase of jumping motion, the number of muscle coordination patterns is more than four and the muscles are more freely organized since there is no contact force with the ground (as analyzed in Section 2.2). To estimate the covariation of joint angles with extracted synergies is a rough task and may cause computational burden. Therefore, we use cubic polynomial interpolation to generate the reference joint angles of flight phase, taking the conditions of the adjacent two jumps. Assume the time of each jump accounting for the moment of the feet touching the ground until the beginning of the next stance phase, which includes stance time t_j $(j = 0, \ldots, (n-1))$ and flight time t_i $(i = n, \ldots, m)$. The joint angle trajectories of flight phase for the Nth jump during continuous jumping can be expressed as

$$\Theta_N^{t_i} = A_N^1 + A_N^2 t_i + A_N^3 t_i^2 + A_N^4 t_i^3 \tag{7}$$

in which

$$\Theta_N^{t_i} = \begin{bmatrix} \left[\theta_N^{t_i}\right]_{\text{hip}} \\ \left[\theta_N^{t_i}\right]_{\text{knee}} \\ \left[\theta_N^{t_i}\right]_{\text{ankle}} \end{bmatrix} \tag{8}$$

$$A_N^l = \begin{bmatrix} \left[a_N^l\right]_{\text{hip}} \\ \left[a_N^l\right]_{\text{knee}} \\ \left[a_N^l\right]_{\text{ankle}} \end{bmatrix}, \quad l = 0, \ldots, 3. \tag{9}$$

All the hip, knee, and ankle joint angles are uniformly expressed in vector $\Theta_N^{t_i}$, and A_N^l is the coefficient of cubic polynomial expression. Take the end state of stance phase for the Nth jump as the initial condition of the flight phase $[\Theta_N^{t_n}, \dot{\Theta}_N^{t_n}]^T$ and the start state of stance phase for the $(N+1)$th jump as the terminating condition $[\Theta_N^{t_m}, \dot{\Theta}_N^{t_m}]^T$:

$$\begin{bmatrix} \Theta_N^{t_n} \\ \dot{\Theta}_N^{t_n} \\ \Theta_N^{t_m} \\ \dot{\Theta}_N^{t_m} \end{bmatrix} = \begin{bmatrix} \Theta_N^{t_{n-1}} \\ \dot{\Theta}_N^{t_{n-1}} \\ \Theta_{N+1}^{t_0} \\ \dot{\Theta}_{N+1}^{t_0} \end{bmatrix} \tag{10}$$

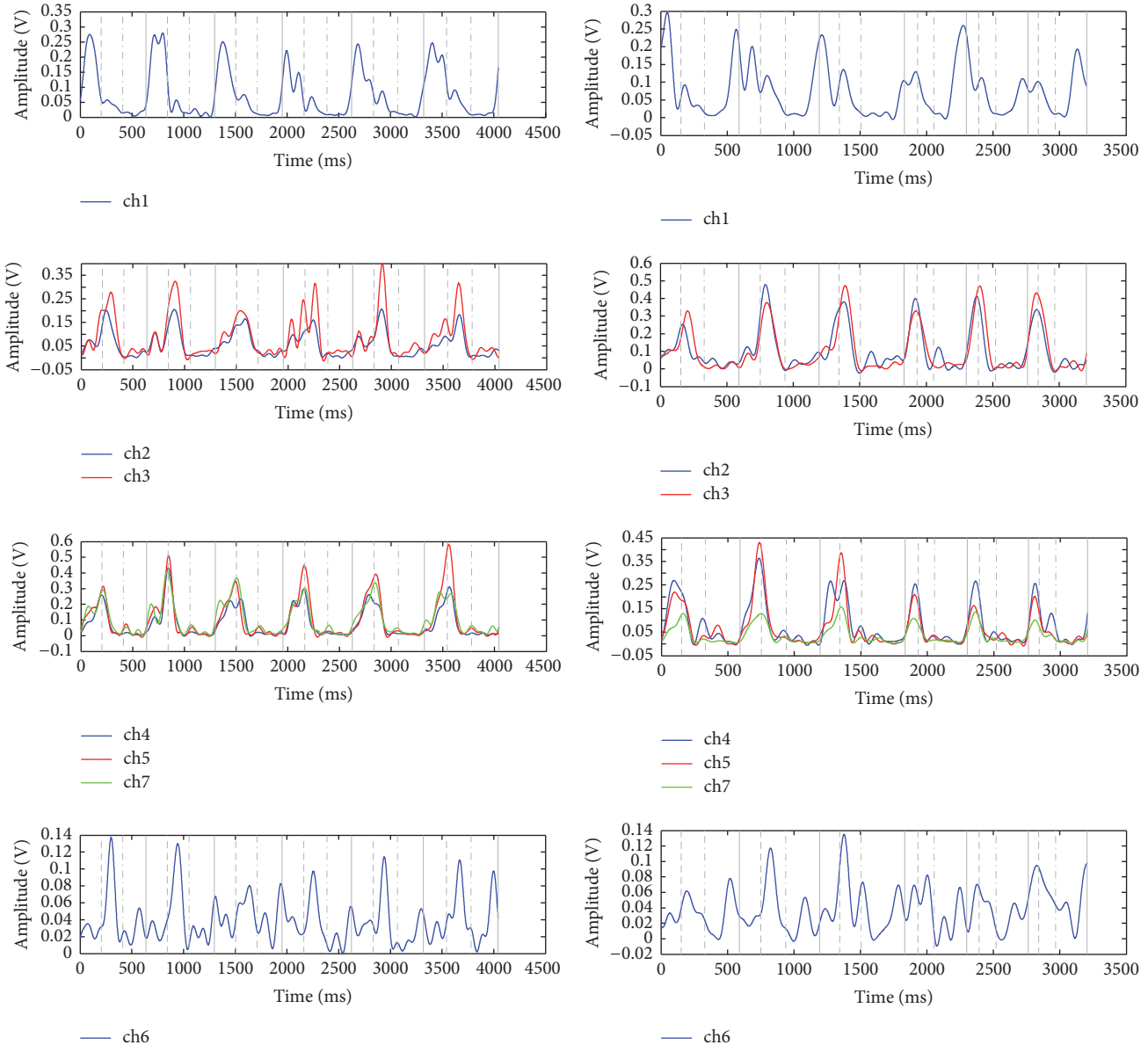
in which $\Theta_N^{t_n}$ and $\dot{\Theta}_N^{t_n}$ are the initial joint angle and angle velocity of flight phase for the Nth jump, while $\Theta_N^{t_m}$ and $\dot{\Theta}_N^{t_m}$ are the initial joint angle and angle velocity, respectively. Substituting (10) into (7), coefficient of cubic polynomial expression A_N^l can be solved.

4.3. Gait Generalization with Takagi-Sugeno Fuzzy Inference System. The proposed fuzzy inference system is based on the well know Takagi-Sugeno fuzzy inference system (TS-FIS). The TS-FIS is described by a set of R_k $(k = 1 \cdots N_k)$ fuzzy rules presented in (11). x_i $(i = 1 \cdots N_i)$ are the inputs of the FIS with N_i dimension input space, and A_i^j $(j = 1 \cdots N_j)$ are linguistic terms, which are numerically defined by membership functions distributed in the universe of discourse for each input x_i. Each output rule y_k is a linear combination of input variables.

$$\text{if } x_1 \text{ is } A_1^j \ldots \text{and} \ldots x_i \text{ is } A_i^j \text{ then } y_k = (x_1, \ldots, x_{Ni}). \tag{11}$$

In our problem, each reference gait has been memorized by one wavelet neural network system, which records the relationship between muscle coactivation coefficient and joint coordinated variation. Based on the available hopping velocity from experiment data, totally six reference gait patterns are chosen and memorized.

The membership functions are shown in Figure 8, in which the desired jumping rhythm is modeled by six



(a) Continuous jumping with uniform velocity

(b) Continuous jumping with variable velocity

FIGURE 9: Joint angles and muscle synergies. ch1, tibialis anterior; ch2, gastrocnemius; ch3, soleus; ch4, vastus medialis; ch5, rectus femoris; ch6, biceps femoris; ch7, gluteus maximus.

fuzzy sets ("VeryFast," "Fast," "MediumFast," "MediumSlow," "Slow," and "VerySlow"). The final desired trajectory of hip, knee, and ankle joint angles is generalized by the TS-FIS and WNN architecture, on the basis of predefined membership functions:

If v_m is VeryFast then $Y = O_1$.

If v_m is Fast then $Y = O_2$.

If v_m is MediumFast then $Y = O_3$.

If v_m is MediumSlow then $Y = O_4$.

If v_m is Slow then $Y = O_5$.

If v_m is VerySlow then $Y = O_6$.

The output Y is carried out in two stages [30].

Firstly, take the extracted coactivation coefficient as the inputs of the wavelets neural network. The coordinating variation of hip, knee, and ankle joint angles $O_k(C)$ can be identified according to (2), indicating each reference gait pattern with corresponding muscle coactivation coefficient.

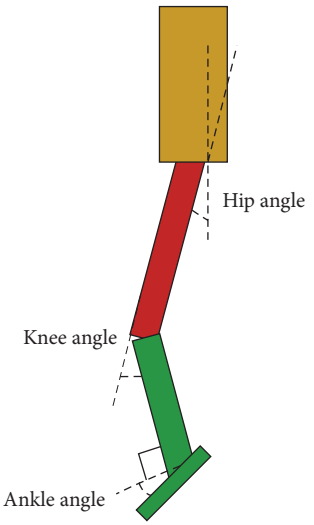


Figure 10: Definition of joint angles in vertical plane.

Secondly, the generalized joint angle Y is estimated using the weighted average of all wavelets neural networks, which is calculated using

$$Y = \sum_{k} \overline{u}_k O_k (C) \tag{12}$$

with $\overline{u}_k$ given by

$$\overline{u}_k = \frac{u_k}{\sum_{k=1}^{N_r} u_k} \tag{13}$$

and u_k is computed with the membership function according to

$$u_k = \mu_1^j \mu_2^j \cdots \mu_{Ni}^j. \tag{14}$$

By using the fuzzification of several outputs of wavelets neural network, it is possible to obtain a global generalization, which allows decreasing the memory size and computing cost using only a small set of identification system. Moreover, it makes the generalization of gait pattern for a variety of velocities possible, with only limited sets of jumping experiment results.

5. Discussion

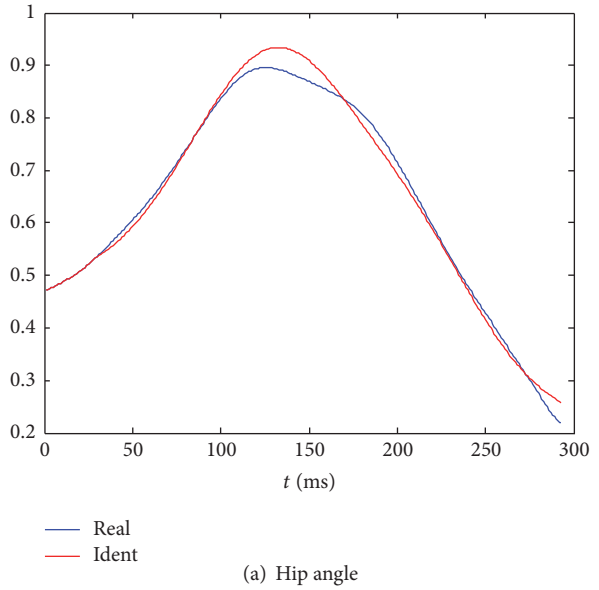
5.1. Muscle Synergies for Stance Phase of Vertical Jumping. If the number of synergies is chosen as four, the simulation results in Table 1 (Section 3.2) show that the VAF of all the trials is above 90%, which indicates that the recorded sEMGs are well reconstructed by the extracted synergies. The number of synergies that can explain the muscle coactivation pattern efficiently was discussed in our previous work in [31]. Take two experimental trials as examples (uniform velocity jumping and variable velocity jumping). The sEMG signals after preprocessing are shown with the same time sequence in Figure 9. The four synergies of stance phase are gastrocnemius, soleus; vastus medialis, rectus femoris, and gluteus maximus; tibialis anterior; and biceps femoris, respectively. The muscles were grouped into one synergy activated with very similar time sequence, while the time series of muscle explosive for each synergy are different from one another. It is noticed that the biceps femoris collected by the first channel sEMG is always activated in front of other muscles and even before the moment of maximum joint compression. This observation coordinates with the conclusion in [32], in which, before the stance phase of jumping, some muscles have already been preactivated in order to decrease the leg impact force with ground.

5.2. Velocity-Adapted Gait Generalization

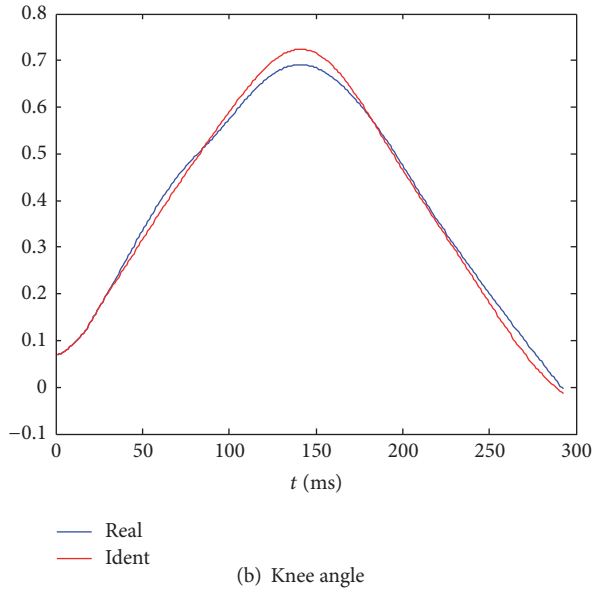
5.2.1. Flexible Vertical Jumping Gait for Stance Phase. The hip, knee, and ankle joint angle defined in bionic leg model are shown in Figure 10. As designed in Section 4.1, a 4-input and 3-output WNN was constructed to predict the coordinating variation of joint angles with muscle synergies for stance phase. Taking the muscle activation coefficients as input signals, the output joint angles of neural network for reference "Gait3" defined in Section 4.3 (jumping with rhythm of 517 ms) are shown in Figure 11, in which blue line expresses the joint angles from motion capture data and red line is the estimated outputs. In order to present the performance of the proposed method more clearly, all the joints were undergoing normalization by the maximum value.

Root mean square error (RMSE) is computed to measure the performance of the wavelet neural network model. The RMSE is defined as follows:

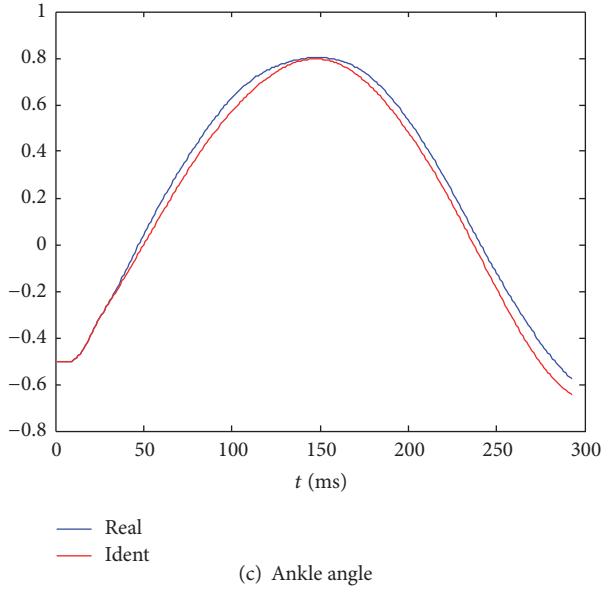
$$\mathrm{RMSE}\left(\left\{\overline{\theta}(i)\right\}, \{\theta(i)\}\right) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{\theta}(i) - \theta(i)\right)^2} \tag{15}$$



(a) Hip angle



(b) Knee angle



(c) Ankle angle

Figure 11: Comparison of estimated joint angles with experimental data for gait pattern 3.

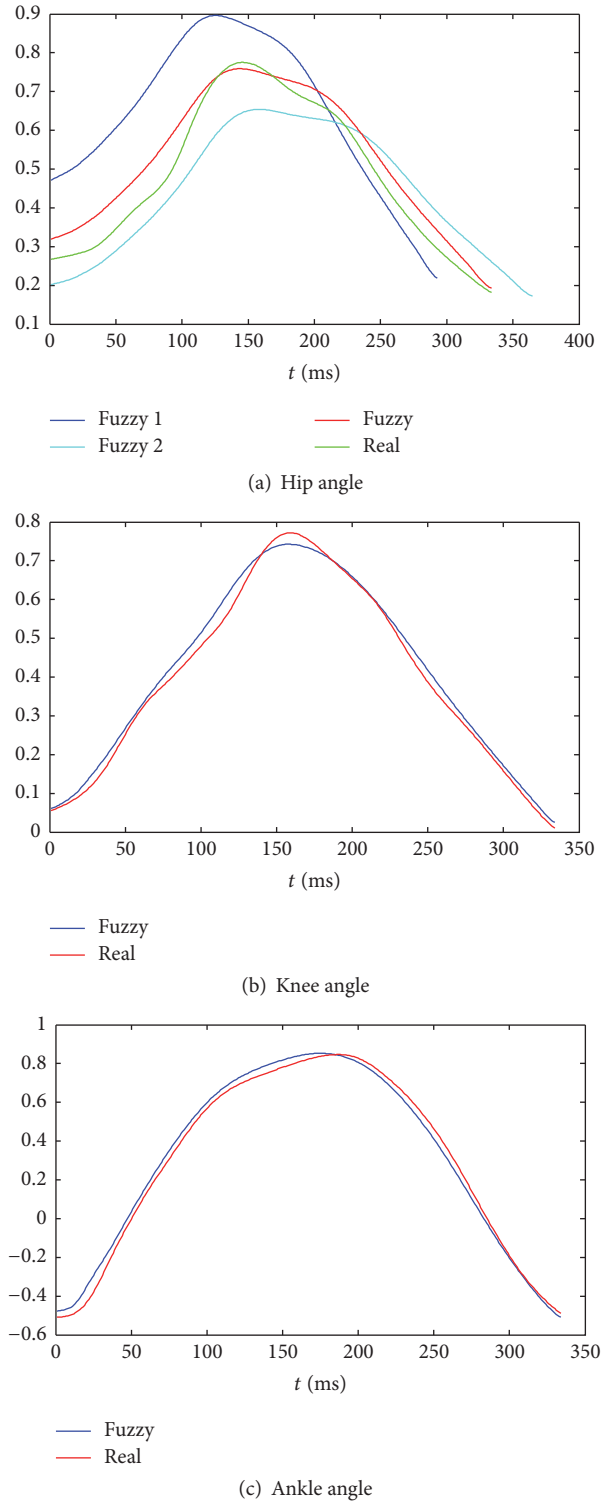
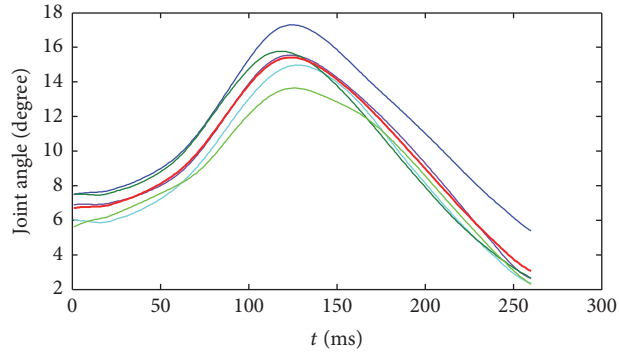
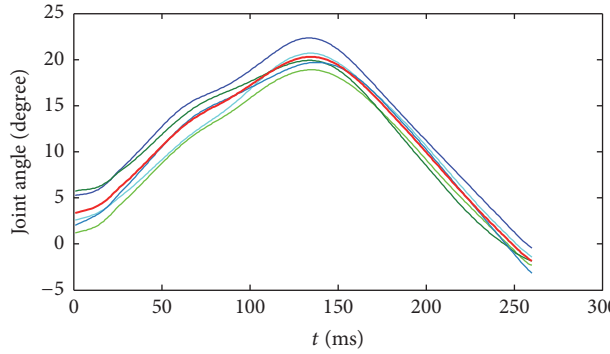


Figure 12: Generalized flexible gait with jumping rhythm of 580 ms.

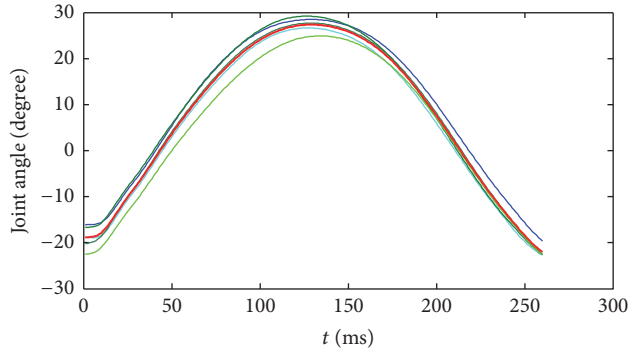
in which $\overline{\theta}(i)$ is the estimated joint angle, $\theta(i)$ is the real joint angel, and N is the length of the sample data. Table 3 gives the RMSEs between the real joint angles from the experimental data and the estimated ones for one subject, which does not have much difference from other five subjects. The results indicate that the proposed WNN model with muscle synergies shows good performance both in accuracy and in robustness for different hopping velocities. Six reference



(a) Hip angle



(b) Knee angle



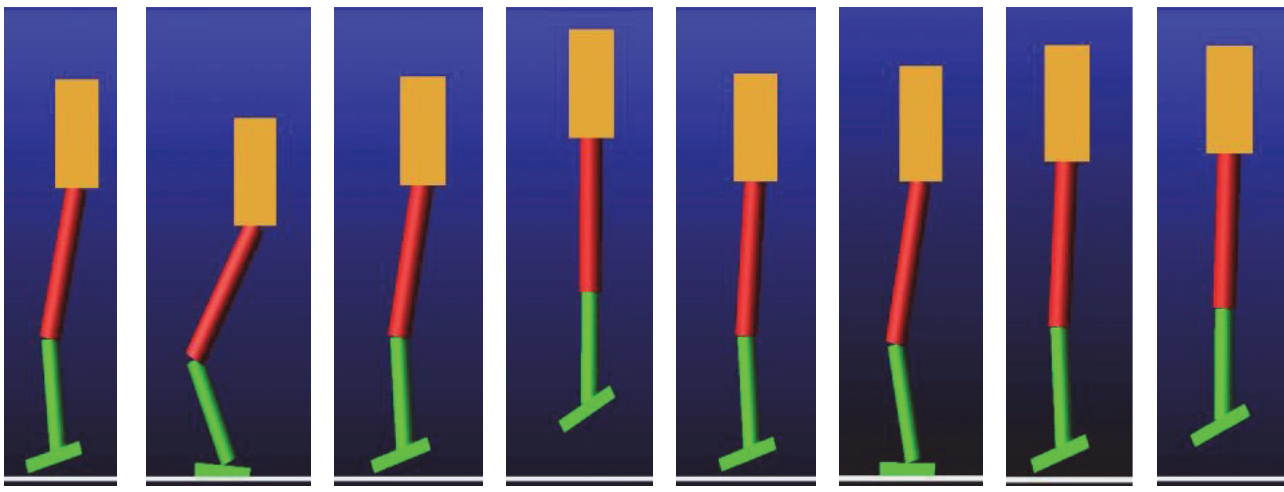
(c) Ankle angle

Figure 13: Experimental joint angles of the same jumping rhythm for stance phase.

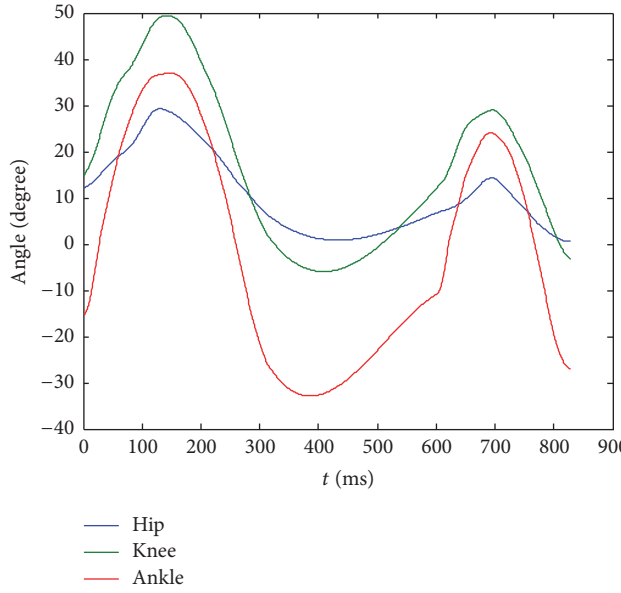
gait patterns of stance phase are memorized in these neural networks.

The fuzzy inference system proposed in Section 4.3 is used to generalize more flexible gait with limited experimental gait pattern. Six triangle membership functions were set according to the different hopping velocity grouped based on the experimental data. Figure 12 presents the generalized gait for stance phase with jumping rhythm of 580 ms after FIS, which refers to the reference gait pattern with cadence of 560 ms and 601 ms, and fuzzy sets of "MediumSlow" and "Slow." The hip angle of the reference gait pattern is especially shown in Figure 12(a). The difference between the real hip angle and the generalized one after FIS is bigger than the other joint angles. Furthermore, we analyze the stance phase gaits of the same hopping velocity obtained from motion capture data. It can be seen from Figure 13 that the hip angles vary more than other joint angles even for the same subject. This may explain why the generalized hip gait is slightly different from the real experimental data.

5.2.2. Velocity-Adapted Continuous Vertical Jumping Gait for Bionic Leg. Next, we evaluate the proposed approach to generate self-adapted continuous vertical jumping for bionic leg with variable velocity. Figure 14 represents the motion sequences and joint angle trajectories with the jumping velocity changing from slow to fast. While the adjacent two jumps from fast to slow jumps are shown in Figure 15. From the results, it should be noticed that, in the case of slow jump



(a) Motion sequences



(b) Joint angle trajectories

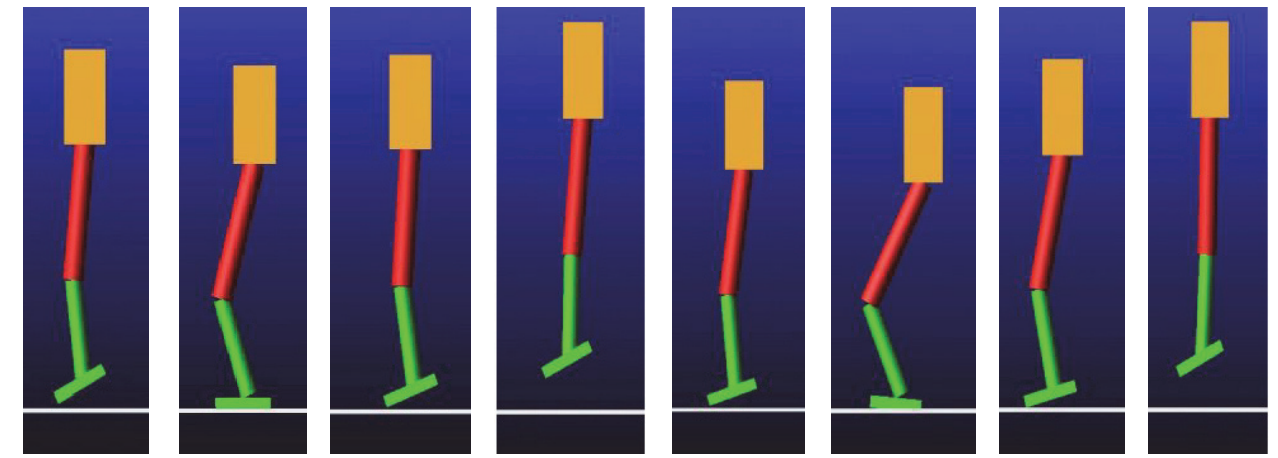
Figure 14: Generalized gait of adjacent two jumps from slow to fast speed.

Table 3: Joint angles estimation RMS error for six reference gait patterns.

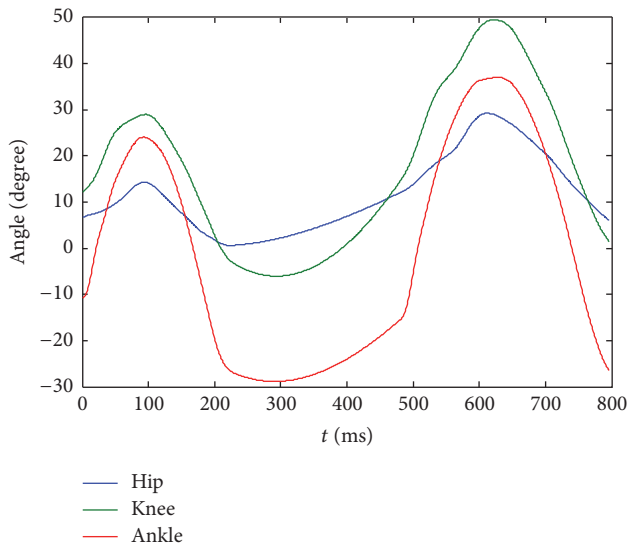
Joint	RMS error (%)					
	Gait 1	Gait 2	Gait 3	Gait 4	Gait 5	Gait 6
Ankle	0.97	3.03	1.90	1.94	4.34	1.41
Knee	1.13	5.17	1.84	1.94	5.07	1.53
Hip	4.62	2.06	4.94	6.75	3.75	4.22

of bionic leg, all of the hip, knee, and ankle angles are more compressed than fast jump. This also follows the regularities observed from the subject jump experiment. The generalized gait is compliant in the condition of changing velocity, either from slow to fast hops or from fast to slow hops.

Figure 16 shows generalized joint angle trajectories of bionic leg for velocity-adapted continuous jump. The results indicate that the proposed approach can generate flexible hopping gait very similar to human motion with limited experimental data. Furthermore, compared with stance phase, the joint angles in the flight phase show smaller fluctuation. That is because the jumping gait of flight phase is generalized with polynomial interpolation simply. With this proposed approach, the self-development of new flexible vertical jumping gait can be achieved only referring to limited



(a) Motion sequences



(b) Joint angle trajectories

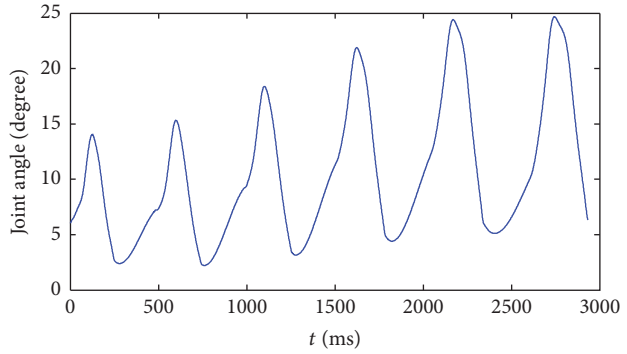
Figure 15: Generalized gait of adjacent two jumps from fast to slow speed.

reference gait patterns, which extremely shortens the training time of estimation model for sEMG-based biomechanical leg control.

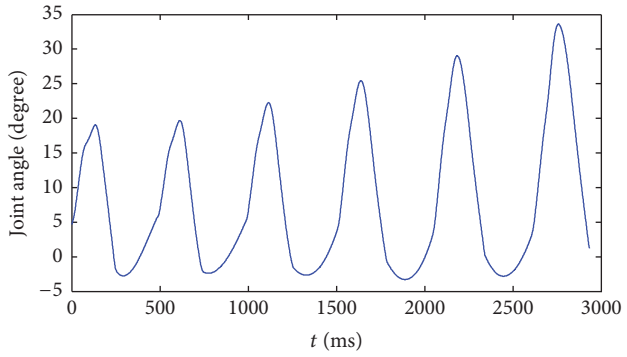
6. Conclusion

Inspired by the hypothesis that CNS modulates muscle synergies to simplify the motor control and learning of coordinating variation of redundant joints, this paper proposed a novel approach for flexible gait generation of hopping motion with sEMG signals. Two questions were analyzed and discussed in the paper, the first one concerning whether the same set of muscle synergies can explain the different phases of jumping movement with various velocities. Based on the analysis of synergies weighing coefficients and variance accounted for (VAF) value, the muscle synergies were extracted for stance phase separately. The second one is about building a model for generating velocity-adapted jumping gait with muscle synergies, in which wavelets neural network is proposed to predict the reference gait pattern, while fuzzy inference system is adopted to merge these reference gaits in order to create more generalized gaits with different jumping rhythm. From the experimental results, the proposed method shows good performance both in accuracy and in robustness for producing continuous flexible jumping gait with different velocities.

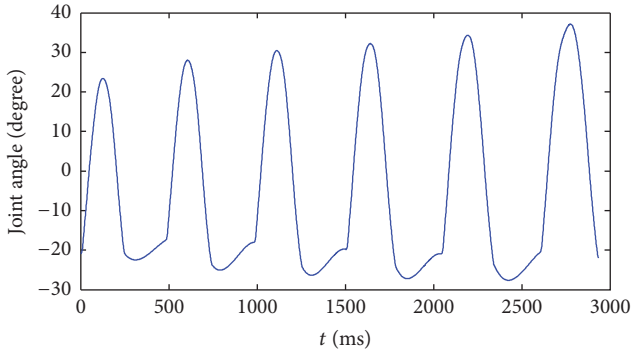
The proposed method can be adopted as the decoder in sEMG-based controls [21] for bionic leg, which decodes muscle activities into intuitive control outputs by training a model on sEMG-related inputs with desired motion gait.



(a) Hip angle trajectory for continuous jumps



(b) Knee angle trajectory for continuous jumps



(c) Ankle angle trajectory for continuous jumps

FIGURE 16: Generalized continuous jumping gait adapted to velocity change.

Once the decoder was trained, it is used in real time to estimate the multijoint angles and map them to output of myoelectric interface. Moreover, linear combinations of synergies are capable of describing complex force and motion patterns in reduced dimensions [33]. Therefore the robust representations of synergies within the control scheme can generate flexible gaits for other complex motions, such as hopping and running with user's intent.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research has been made possible by "111 Project" (Grant no. B13044), Natural Science Basic Research Plan of Shaanxi Province (Grant no. 2016JQ6009), and National Science Foundation of China (Grant no. 61603302).

References

[1] M. Pang, S. Guo, Q. Huang, H. Ishihara, and H. Hirata, "Electromyography-based quantitative representation method for upper-limb elbow joint angle in sagittal plane," *Journal of Medical and Biological Engineering*, vol. 35, no. 2, pp. 165–177, 2015.

[2] D. P. Yang, L. Jiang, Q. Huang, R. Liu, and H. Liu, "Experimental study of an EMG-controlled 5-DOF anthropomorphic prosthetic hand for motion restoration," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 76, no. 3-4, pp. 427–441, 2014.

[3] M. Ison and P. Artemiadis, "Proportional myoelectric control of robots: muscle synergy development drives performance enhancement, retainment, and generalization," *IEEE Transactions on Robotics*, vol. 31, no. 2, pp. 259–268, 2015.

[4] M. Hakonen, H. Piitulainen, and A. Visala, "Current state of digital signal processing in myoelectric interfaces and related applications," *Biomedical Signal Processing and Control*, vol. 18, pp. 334–359, 2015.

[5] A. d'Avella, P. Saltiel, and E. Bizzi, "Combinations of muscle synergies in the construction of a natural motor behavior," *Nature Neuroscience*, vol. 6, no. 3, pp. 300–308, 2003.

[6] A. d'Avella and E. Bizzi, "Shared and specific muscle synergies in natural motor behaviors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 8, pp. 3076–3081, 2005.

[7] N. Kuppuswamy and C. M. Harris, "Do muscle synergies reduce the dimensionality of behavior?" *Frontiers in Computational Neuroscience*, vol. 8, article no. 63, 2014.

[8] M. Berniker, A. Jarc, E. Bizzi, and M. C. Tresch, "Simplified and effective motor control based on muscle synergies to exploit musculoskeletal dynamics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 18, pp. 7601–7606, 2009.

[9] C. Alessandro, I. Delis, F. Nori, S. Panzeri, and B. Berret, "Muscle synergies in neuroscience and robotics: from input-space to task-space perspectives," *Frontiers in Computational Neuroscience*, vol. 7, no. 43, pp. 1–16, 2013.

[10] G. Cappellini, Y. P. Ivanenko, R. E. Poppele, and F. Lacquaniti, "Motor patterns in human walking and running," *Journal of Neurophysiology*, vol. 95, no. 6, pp. 3426–3437, 2006.

[11] C. C. Raasch and F. E. Zajac, "Locomotor strategy for pedaling: muscle groups and biomechanical functions," *Journal of Neurophysiology*, vol. 82, no. 2, pp. 515–525, 1999.

[12] S. A. Chvatal, G. Torres-Oviedo, S. A. Safavynia, and L. H. Ting, "Common muscle synergies for control of center of mass and force in nonstepping and stepping postural behaviors," *Journal of Neurophysiology*, vol. 106, no. 2, pp. 999–1015, 2011.

[13] J. L. Allen and R. R. Neptune, "Three-dimensional modular control of human walking," *Journal of Biomechanics*, vol. 45, no. 12, pp. 2157–2163, 2012.

[14] D. J. Clark, L. H. Ting, F. E. Zajac, R. R. Neptune, and S. A. Kautz, "Merging of healthy motor modules predicts reduced

locomotor performance and muscle coordination complexity post-stroke," *Journal of Neurophysiology*, vol. 103, no. 2, pp. 844–857, 2010.

[15] E. J. Weiss and M. Flanders, "Muscular and postural synergies of the human hand," *Journal of Neurophysiology*, vol. 92, no. 1, pp. 523–535, 2004.

[16] P. K. Artemiadis and K. J. Kyriakopoulos, "EMG-based control of a robot arm using low-dimensional embeddings," *IEEE Transactions on Robotics*, vol. 26, no. 2, pp. 393–398, 2010.

[17] S. Muceli, N. Jiang, and D. Farina, "Extracting signals robust to electrode number and shift for online simultaneous and proportional myoelectric control by factorization algorithms," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 3, pp. 623–633, 2014.

[18] J. M. Hahne, F. Biebmann, N. Jiang et al., "Linear and nonlinear regression techniques for simultaneous and proportional myoelectric control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 2, pp. 269–279, 2014.

[19] H. Kalani, S. Moghimi, and A. Akbarzadeh, "SEMG-based prediction of masticatory kinematics in rhythmic clenching movements," *Biomedical Signal Processing and Control*, vol. 20, pp. 24–34, 2015.

[20] S. Muceli and D. Farina, "Simultaneous and proportional estimation of hand kinematics from EMG during mirrored movements at multiple degrees-of-freedom," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20, no. 3, pp. 371–378, 2012.

[21] M. Ison and P. Artemiadis, "The role of muscle synergies in myoelectric control: Trends and challenges for simultaneous multifunction control," *Journal of Neural Engineering*, vol. 11, no. 5, Article ID 051001, 2014.

[22] J. L. McKay and L. H. Ting, "Optimization of muscle activity for task-level goals predicts complex changes in limb forces across biomechanical contexts," *PLoS Computational Biology*, vol. 8, no. 4, Article ID e1002465, 2012.

[23] J. J. Kutch and F. J. Valero-Cuevas, "Challenges and new approaches to proving the existence of muscle synergies of neural origin," *PLoS Computational Biology*, vol. 8, no. 5, Article ID e1002434, 2012.

[24] M. C. Tresch, V. C. K. Cheung, and A. d'Avella, "Matrix factorization algorithms for the identification of muscle synergies: evaluation on simulated and experimental data sets," *Journal of Neurophysiology*, vol. 95, no. 4, pp. 2199–2212, 2006.

[25] C. Castellini and P. Van Der Smagt, "Evidence of muscle synergies during human grasping," *Biological Cybernetics*, vol. 107, no. 2, pp. 233–245, 2013.

[26] A. B. Ajiboye and R. F. Weir, "Muscle synergies as a predictive framework for the EMG patterns of new hand postures," *Journal of Neural Engineering*, vol. 6, no. 3, Article ID 036004, 2009.

[27] X. Chen, X. Zhu, and D. Zhang, "A discriminant bispectrum feature for surface electromyogram signal classification," *Medical Engineering and Physics*, vol. 32, no. 2, pp. 126–135, 2010.

[28] J.-U. Chu, I. Moon, S.-K. Kim, and M.-S. Mun, "Control of multifunction myoelectric hand using a real-time EMG pattern recognition," in *Proceedings of the IEEE IRS/RSJ International Conference on Intelligent Robots and Systems, IROS 2005*, pp. 3957–3962, can, August 2005.

[29] M. Russo, M. D'Andola, A. Portone, F. Lacquaniti, and A. d'Avella, "Dimensionality of joint torques and muscle patterns for reaching," *Frontiers in Computational Neuroscience*, vol. 8, article no. 24, 2014.

[30] C. Sabourin, W. Yu, and K. Madani, "Gait pattern based on CMAC neural network for robotic applications," *Neural Processing Letters*, vol. 38, no. 2, pp. 261–279, 2013.

[31] W. Yu, Y. Feng, W. Liang, R. Wang, and K. Madani, "Estimate the kinematics with EMG signal using fuzzy wavelet neural network for biomechanical leg application," in *Advances in Swarm Intelligence*, vol. 9713 of *Lecture Notes in Computer Science*, pp. 132–140, Springer International Publishing, 2016.

[32] S. Rapoport, J. Mizrahi, E. Kimmel, O. Verbitsky, and E. Isakov, "Constant and variable stiffness and damping of the leg joints in human hopping," *Journal of Biomechanical Engineering*, vol. 125, no. 4, pp. 507–514, 2003.

[33] A. de Rugy, G. E. Loeb, and T. J. Carroll, "Are muscle synergies useful for neural control?" *Frontiers in Computational Neuroscience*, 2013.

Hindawi
Journal of Sensors
Volume 2017, Article ID 6043069, 18 pages
https://doi.org/10.1155/2017/6043069

Research Article

An IoT Platform for Epilepsy Monitoring and Supervising

Paula M. Vergara,[1] Enrique de la Cal,[1] José R. Villar,[1] Víctor M. González,[2] and Javier Sedano[3]

[1]*Computer Science Department, University of Oviedo, Campus de Viesques, s/n, 33204 Gijón, Spain*
[2]*DIEECS, University of Oviedo, Campus de Viesques, s/n, 33204 Gijón, Spain*
[3]*Instituto Tecnológico de Castilla y León, Polg. Ind. Villalonquejar, López Bravo 70, 09001 Burgos, Spain*

Correspondence should be addressed to José R. Villar; villarjose@uniovi.es

Received 9 March 2017; Revised 29 May 2017; Accepted 6 June 2017; Published 27 July 2017

Academic Editor: Guiyun Tian

Epilepsy is a chronic neurological disorder with several different types of seizures, some of them characterized by involuntary recurrent convulsions, which have a great impact on the everyday life of the patients. Several solutions have been proposed in the literature to detect this type of seizures and to monitor the patient; however, these approaches lack in ergonomic issues and in the suitable integration with the health system. This research makes an in-depth analysis of the main factors that an epileptic detection and monitoring tool should accomplish. Furthermore, we introduce the architecture for a specific epilepsy detection and monitoring platform, fulfilling these factors. Special attention has been given to the part of the system the patient should wear, providing details of this part of the platform. Finally, a partial implementation has been deployed and several tests have been proposed and carried out in order to make some design decisions.

1. Introduction

Epilepsy is a chronic neurological disorder characterized by involuntary recurrent convulsions [1]. There are about 65 million people affected all around the world, with a high and dramatic impact not only on the patient's quality of life, but also on the professional development and social behaviour; the health system budget is highly affected as well.

The illness anamnesis improves with the existing platforms for patient monitoring and weblogs. The main part of these platforms has been developed for two different and most frequent kinds of epilepsy crisis: the generalized tonic-clonic seizures and the typical absence seizures [2]. In these two cases, the detection of a seizure can be efficiently faced using wearable sensors (WDs) including a triaxial accelerometer (ACM) and/or a heart rate (HR) sensor: the former type detection has been reported in [3], the latter one has been characterized in [4], and an HR-based detection system has been proposed in [5].

Another main aspect of the anamnesis process is where the data is gathered. The main part of the literature deals with constrained spaces, that is, research laboratories or hospital rooms [6], or even the patient's house [7], but without considering the normal everyday life [8, 9]. We claim that the data should be gathered in everyday life, allowing the patient to freely decide what to do and how to do it. This is important because, firstly, the data is gathered from normal activities performed before and after a seizure, and secondly, the analysis and procedures should adapt to this unconstrained world, making the whole detection process much more difficult.

A careful in-depth analysis of the seminal papers concerning epilepsy monitoring platforms [10] and Mobile Cloud Computing (MCC) [11–13] let us conclude that the current available platform, either in the scientific literature or in the market, lacks several main features that are not comprehensibly integrated. These include, among others, the instantaneous data source problem, the ergonomic aspects in the design, the deployment cost of the solution, and the energy efficiency of the approach.

This study aims to solve some of these limitations; to do so, a solution is proposed and an experimentation stage has been performed in order to extract the suitable conclusions for the epilepsy monitoring platforms. In the next section, the

most relevant contributions in the literature are analyzed and criticized, paying special attention to the published platforms; the main concerns that remain unsolved are included as well. Section 3 is devoted to explaining an architecture that solves the main concerns found in the literature related to the design of epilepsy monitoring platforms. This section also describes the MCC and discusses some design parameters. Finally, an experimentation stage has been performed and the results are included and discussed in Section 4, which allows us to produce the final conclusions. For the sake of readability, Abbreviations includes the most relevant acronyms used within the text.

2. Related Work

Several different eHealth platforms have been released and reported in the literature for the detection and/or monitoring of illnesses in real time [14], even using WD and/or Body Sensor Networks (BSN) [15, 16]. Typically, the WD are proposed for data gathering, measuring biomedical variables, or obtaining feedbacks from the users, either performing local MCC processing [17, 18] or requesting Cloud Computing (CC) services [19–21]. Usually, the CC services are responsible for processing and storing the sampled data from sensors, as well as the model learning, and those computational greedy tasks. Additionally, the CC services also provide the presentation layer, in terms of user alarms to the patients or medical staff, notifications to the patient's relatives, or even performing graphics and data analytics for further studies.

Examples of such platforms include CoCaMaal, ROCHAS, and AACMPE. CoCaMaal, the acronym for A cloud-oriented context-aware middleware in ambient assisted living [21], is specialized in the patient monitoring and in the event control, either notifications or accidents. This platform restricts itself to fully controlled environments as long as it suggests the deployment of BSN according to the patient's conditions. Variables such as electroencephalogram (EEG), electrocardiogram (ECG), ACM, HR, and blood pressure are considered to be placed on each patient; therefore, the ergonomic aspects of this solution need checking.

A second interesting platform is ROCHAS (Robotics and Cloud-assisted Healthcare System for Empty Nester) [22], which proposes the monitoring of handicapped patients in their own home, allowing them to live as independently as possible by means of an assistance robot. Similarly, an assistant platform for elderly people was proposed in a series of studies in [19, 20, 23], where open software platforms are analyzed to work together.

AACMPE, short term for Allergy and Asthma Care in the Mobile Phone Era [24], controls the allergies and asthma evolution of the patients using MCC to determine several variables: the peak exhalation flow and the peak nasal inhalation flow and some breath parameters and sounds, among others. The Chinese CMTHC project, Children's medical treatment and healthcare system [25], has been reported to monitor unhealthy children by means of web logs to be completed with sensor data, like body temperature, HR, and so on. A further step has been proposed in AIWAC, affective interaction through wearable computing and cloud technology [26], which analyzes the affective needs of the patients based on the measurements obtained from a WD. Other platforms that propose open solutions for monitoring illnesses introduce BSN in a more abstract way. Further work on adapting them for epilepsy detection is needed [27, 28].

Concerning the epilepsy detection and monitoring, several approaches have been reported in the literature. The main data source to do so is the EEG, measuring the electrical activity of the brain to detect the epileptic seizures. Advances in some issues have been published, like modelling the recorded signals [29, 30] or the design of portable EEG devices to deploy such models.

Some examples of these epilepsy specific platforms include EpiCare, Sareen, or Bajwa. In EpiCare, a home care platform based on Mobile Cloud Computing to assist epilepsy diagnosis [7], portable EEG are used to detect the seizures in controlled environments. Sareen gathers data from EEC, using MCC and CC for storage and for sending notifications to the relatives and medical staff, including location information [31, 32]. The main difference between Sareen and Bajwa is that the latter proposes CC only.

EEG aside, different biometric signals have been proposed for detecting the epileptic seizures [10, 33]: gyroscopes, magnetometers, implanted advisory system, electromyography, ECG, ACM, video detection systems, mattress sensor, and audio classification. BSN placed on the body together with MCC has been proposed for epilepsy detection in [8]; more specifically, the BSN includes an ACM cap, a wrist band including a temperature sensor and an ACM, a moisture sensor, and microphone. The MCC layer analyzes the data and detects the seizures; the relevant data is stored and transmitted to the medical staff for further analysis.

Furthermore, solutions making use of CC services have been also reported, mainly for the storage and modelling of the gathered data. For instance, not all the platforms store the data stream; on the contrary, the majority of them process the data as it comes in order to generate the alarms and delete them afterwards: these solutions rely on previous research stages that would have been performed to obtain the deployed models [6, 8, 9, 34–38]. On the other hand, there are solutions where the gathered data is stored for data analytics [39, 40] or even for future use [41–44]. Finally, complete CC solutions have been also tackled in the literature, including not only the data storage and visualization but also the modelling and classification of the current state of affairs [7, 31, 32, 45–48].

Soon after the study of Schulc et al. [36], ACM was proposed for epileptic seizure detection; since then, plenty of studies concerning with the detection of this type of seizure have been published, focusing only on the machine learning issues. For instance, a wrist band including an ACM connected to a Smartphone was proposed for MCC-based seizure detection [3, 6, 49]; no further connection with CC services was considered. ACM, gyroscopes, and magnetometers were proposed in [9] as the BSN, taking advantage of the local Smartphone for analyzing the data and sending e-mails to the medical staff with the patient position. Similar works were presented in [47, 48]. LabView has been used for developing a solution as well [35] by using an Arduino ACM sensor to generate alarms that are transmitted to the medical service.

SmartMonitor® devices have been used for detection of limb shaking, sending the alarms to a website linking the WD with a Smartphone [44]. ECG for detecting epileptic seizures while sleeping has been reported in [39], while EEG hats have been also effectively used in seizure detection [41, 50]; however their ergonomic characteristics make them difficult to use. Further studies in this context of epileptic seizure detection include BSN sending information to a computer in controlled environments [42, 43]; ACM, temperature, and skin humidity data gathered to computers [40]; ACM and HR [5]; or the use of thresholds as a detection method when measuring the HR linked to a Smartphone [38].

Finally, plenty of apps have been published in the corresponding markets for the detection using either the Smartphone sensors or external sensors, EpDetec [51] and MyEpiPal [34], or for the web logging, facilitating the way a patient records daily information concerning her/his epileptic events, medication, and news, My Epilepsy Diary [52], EpiDiary [53], Epilepsy Society [37], and Epi & Me [54]. A reader who is interested should refer to [55] for a performance comparison of epilepsy related apps.

2.1. Remarkable Factors. After this thorough analysis of the literature, we have found out that several remarkable factors have not received enough attention from the research community concerning epilepsy seizure detection and patient monitoring platforms.

Real Time Response. The detection of seizures and the alarm notification should work almost immediately. That means that once a seizure sets in, in the smallest time lapse the alarms should notify the relatives and the health services. A balance between stand-alone solutions, that is, everything computed in the mobile device, and the autonomy and the battery life should be achieved: the higher the computation, the shorter the battery life and thus the autonomy. Therefore, some services must be declared as essential, while others might be postponed, storing the information in a local database. Furthermore, the features and capabilities for each mobile device must be properly introduced in the system, so a suitable ordering of the services can be generated.

Ergonomic Issues. The solutions should be easy to wear and unobtrusive: the better the wearable conditions, the smaller the chances of finding a suitable excuse for not using the solution. Ergonomic issues also include factors like energy efficiency and long battery life: the lack of them forces frequent charging cycles of the devices and some other annoyances. Besides, some examples of devices without lacking in ergonomic issues are found in the main part of the solutions; they are efficient for their purpose, but their use is uncomfortable though: using sensing caps [7, 8, 31, 32], wrongly sized WD [36, 41], too many WDs [8], and so on.

Deployment Cost of the Platform. As for the ergonomic issues, no further study has been considered about the cost of the solution: are the WDs affordable? How much computation would be needed? Can the solution be delivered as a general service? Are tuning stages required? Is specific training of the users needed? Are the solutions economically feasible? How are they contributing to the health system and at what cost? Unfortunately, these questions have been sparsely answered in the published solutions.

CC Guaranteeing the Service Delivery. CC services should be responsible for making all the services available to the users. In spite of the distribution of the computational tasks among all the available hardware, this CC layer should solve any task that no other element within the infrastructure can afford. Besides, complex tasks, such as extension mechanisms that include on-demand model learning, need to be addressed in this layer as well: learning algorithms need relatively large datasets and high computational resources that by no means can be faced on, for instance, mobile devices without penalizing the battery life. Therefore, according to the current scenario and user, there would be the need to store data from the sensors within the CC layer. However, there could be some learning tasks that can be distributed as well, such as those related to active learning stages that might be employed.

All of these considerations lead back to web service ontology development [56–58], where the different services, and tasks, were completely defined as well as the relationships among them. This knowledge representation allows a mediator, or scheduler, to set out a plan and allocate the tasks and services on the different nodes. Extending such ontologies with the actuators, the set of computational nodes and devices, and their features would allow scheduling the tasks for each scenario. Furthermore, new services, and tasks, that extend the functionality of the platform can also be described in ontology terms, so their use can be hot-deployed without stopping and reconfiguring the platform.

MCC Integration. When required, the Smartphone can take the control of the sequence of tasks to perform. This would happen mainly when no Wi-Fi network is available for sending the data; actually, this MCC might help in balancing the battery life and the networking load, reducing the required 4G data limit and, thus, the fares to pay. With MCC we understand a combination of CC services together with mobile computing interconnected by means of wireless networks [59–63]. However, we move one step forward in the mobile computing part, introducing the concepts suggested in [12, 64]: we propose the use of hybrid mobile applications, where several services provided by the CC can also be dynamically allocated and performed by the mobile computation, provided there are resources available. In other words, the mobile device can perform as a local cloudlet.

Configurable Services within the Apps. Instead of running all the possible stack of services, it should be desirable that the app adapts to the patient as much as possible, disabling those services that are no longer needed while, at the same time, including new added value services, such as web logging capabilities and affective computing issues. Again, ontologies can help in deploying these issues.

A comparison between the different mentioned platforms solutions is shown Table 1, paying attention to those factors that have been considered relevant within the

Table 1: Main epilepsy detection platforms and solutions published so far. A description with some concerns with respect to the relevant factors is detailed. We suggest [10, 55] for further reading.

Ref.	Description
[41]	An EEG sensory device linked to a Smartphone, which performs some processing of the data, is used together with CC services. However, this is a very short communication and does not provide details of the solution.
[6]	This is mostly a research study focused only on the detection of tonic-clonic epileptic seizures. A WD without wireless communication stored data and some machine learning methods were performed for obtaining offline models.
[7]	In EpiCare, the Android app runs on the Smartphone device, gathering data from a EEG electrodes cap. This hardware makes this solution very efficient but rather uncomfortable. The main thing is that the project is focused on SUDEP; thus it is feasible to have a cap while sleeping. A mixture of CC and MCC solution is proposed, suggesting that an intelligent task delivering and allocation should be performed but without proposing any viable technique.
[34]	MyEpiPal is an app that monitors the patient, simplifies the intercommunication between the caregiver and the patient, and allows the self-management. This means that although it makes use of the sensory within the Smartphone, the main goal is to give support to the patient in everyday life. It is not an epilepsy detection platform itself, although the measurements can help in the prediction of the occurrence of a seizure, which is the reason that it is included in this comparison.
[39]	This research details the design of an ad hoc epilepsy detection ECG wireless intelligent sensor, including several detection algorithms, linked to a local computer connected to a network. Several relevant factors were analyzed, the ergonomic issues and the battery life among them. The WD communicates with the local computer in order to deliver alarms, to receive configuration commands, or to start/stop HR recording to be downloaded to the computer. A very detailed explanation of the requirements and of the hardware decisions is included.
[31, 32]	An EEG cap linked to a Smartphone is arranged to send the gathered data to CC services. Whenever a seizure is detected on the cloud, GPS locations are shared through the notification system. Neither the ergonomic issues nor the battery life and autonomy and the economic costs of the platform have been analyzed.
[37]	The UK Epilepsy Society published an app for Android and iPhone as a Web-log of the seizures, medication monitoring, and so on. This solution is a standalone solution, and the data gathered with this tool is not shared with any health service.
[38]	A standalone mobile solution is proposed using a commercial Smartwatch (MIO Alpha) linked to an Android phone. The app is responsible for gathering and processing the data, as well as the modules for the epileptic seizure detection. The solution does not consider intelligent modules, just some thresholds; this technique is by no means valid in this type of detection.

literature: deployment cost of the platform, real time response, ergonomic issues, MCC integration, multiple services within the Apps, and, finally, CC online modelling.

As can be seen, the problem of designing a platform for the specific problem of epileptic seizure detection and monitoring has not been completely addressed in the literature. Although some general design principles are still valid, some of them need special care to propose a final solution. The next section deals with the decisions for the design of an epileptic monitoring and supervising platform, providing an abstract architecture and the description of a prototype for the evaluation of some of the above-mentioned factors. These decisions include ontology driven tasks, dynamic data gathering and modelling, and the integration of MCC and CC.

3. A Platform for Monitoring and Supervising Epileptic Patients

This study proposes a platform for monitoring and supervising epileptic patients, focused on the two main epilepsy types: the focal myoclonic and the epileptic absence seizures. This platform aims to solve each of the most relevant factors seen in the previous section, providing some extensibility to future developments. This section is devoted to describing this platform as well as the developed prototype. The next subsection introduces some considered concept decisions and requirements, while Section 3.2 describes the abstract architecture. The remaining part of the section, Section 3.3, focuses on detailing the developed prototype.

3.1. Design Decisions and Requirements. This study proposes the use of noninvasive WD such as a sensory bracelet plus a Smartphone to allow the patients to carry on with their own life, performing their everyday activities without the need for specific clothes or garments. In this study, a pair of a WD linked by means of bluetooth 4.0 Low Energy to a Smartphone is called a patient's kit (PK). This solution enhances the ergonomic issues of the solution while favouring the patients to continue using the system. The WD should include ACM and HR sensors in order to detect the two focused types of epileptic seizures. As mentioned in the related work section, there are several studies for detecting the focal myoclonic type of seizure, while further study is needed to detect the second focused type. In order to detect the focal myoclonic seizures, the research published in [3, 65] is proposed. Further study is needed to solve the second type of seizures. The selection of those techniques has been taken due to two main

reasons: (i) the obtained results and (ii) the simplicity of the used models. This simplicity would eventually allow them to be implement so they can run on any available platform, the Smartphones among them.

Taking advantage of the computational capacities of mid-range Smartphones, this study proposes that, besides data gathering and processing and perhaps simple thresholding, the MCC services can be extended to incorporate local model evaluation. To do so, incremental deployment/delivery of trained/tuned models into the MCC kernel would allow continuing the monitoring, providing real time response even when the CC is totally unavailable through Wi-Fi networks. Nevertheless, with the aim of extending the battery life as much as possible, a mixture of MCC and CC together with a suitable balancing algorithm to decide where the decisions or calculations are shall be accomplished. The study of balancing MCC computation, decreasing the communication acts, and the CC computation, decreasing the computation amount in the Smartphone, is one of the contributions of this work, as will be shown later.

Besides, ontology driven tasks and dynamic data gathering personalization are required in order to extend the system. With ontology driven tasks we refer to designing and developing ontologies that describe every single task: from sampling to alarming and notifying. The concept of ontology driven tasks facilitates extending the system, so new procedures can be easily conceptually developed and distributed and deployed in any of the available computation layers. On the other hand, dynamic data gathering personalization refers to marks for which patients data should be gathered. For instance, developing new services, say, detecting a new type of seizure, needs specific data to be gathered for further processing and analysis. However, it is impossible to gather data from every patient as long as the amount of data to store grows unbounded. In addition, it is better to gather data only for patients that might perform the desired event to detect or identify. Actually, the development of ontology driven tasks allows the dynamic data gathering to be implemented as the latter can be viewed as a new task devoted for a specific group of patients.

In addition, the CC services should be performed on low-cost servers that can even be deployed in different public sites, in outpatient clinics, for instance, shifting the computation resources to the endpoints. These servers on the far edge must be federated to avoid data losses and to enhance the performance of the whole system. Furthermore, local nodes with available unused computing resources, personal computers or even personal servers, can be designated to become part of the solution, enhancing the overall computational capacities of each installation while keeping the low-cost profile. Clearly, for this latter case, it might be advisable to sign a special commitment between the user and the health system for accomplishing data and privacy regulations.

Finally, some decisions should be taken regarding with the data analysis and data monitoring and with the system extension capabilities. On the one hand, the Exploiting and Data Analytics module is needed. This module tackles the monitoring and tracking of the patients, showing the main facts to the medical staff. This module also includes human-machine interfaces; the main part of these interfaces should be light clients, HTML5/REST clients or similar, based on Bootstrap technology, so they keep the responsiveness. On the other hand, an extension mechanism must be provided. For instance, the data analysis should allow the medical unit staff to perform a high level of machine learning experimentation that might lead to models for detecting or enhancing different types of epileptic seizures. Intelligent interfaces, similar to KEEL [66] or WEKA [67], should define the offline tasks, their outcomes, and reports. In addition, these interfaces should also allow defining new sequences in the ontology driven system outlined before. For sure, unless the medical unit staffs incorporate multidisciplinary teams, these extension mechanisms must be kept simple and elementary.

3.2. The System Abstract Architecture. Let us call a scenario a concrete specification of the computing devices that are available for solving all the calculations needed to detect and monitor a certain patient. Let us assume that a complete ontology of services, tasks, computing devices, and scenarios is obtained; this ontology could be based on that presented in [56]. Let us assume as well that this ontology is fulfilled for the algorithms, tasks, and services involved in this project. Let us also assume that the algorithms and tasks have been properly implemented to run either on an app, on a server, or on both and that all these varieties are reflected in the ontology data.

According to [56], any sequence of tasks can be completely represented in the ontology, and a mediator can locate each of them on a concrete machine. In this approach, we assume a mediator that assigns the tasks to be executed and where they are to be carried out for each case. Therefore, for each patient and scenario, a sequence of tasks can be planned and allocated. In other words, we are able to define a specific planification and task allocation for each patient and scenario; both of them are made explicit in ontology basis; and every single computing device, including the Smartphones, has access to this information.

The proposed abstract architecture is depicted in Figure 1 as a solution for this very specific epilepsy monitoring and supervising platform. A PK is conformed with a WD plus a Smartphone; the Smartphone includes a complete app that, together with the ontology and the current scheduling, performs the sensor sampling, and so on. Whenever available, Wi-Fi networks are used to send data bunches to be stored in the health service. Nevertheless, notifications and alarms, when generated, should be delivered using the available connectivity. In addition, some spaces can host specific hardware performing as a federated CC server. In any case in which Wi-Fi connectivity is available, the Smartphone must delegate on these systems in order to alleviate the computational requirements for the sake of extending the battery life.

Besides, the health system services, both the CC services and the data storage, perform all the data storage and the computation that remain unsolved in the system, including detection of seizures, alarm generation, and notifications, as well as the reporting for the medical staff. The central CC services, those reflected in the DMZ zone, cope with

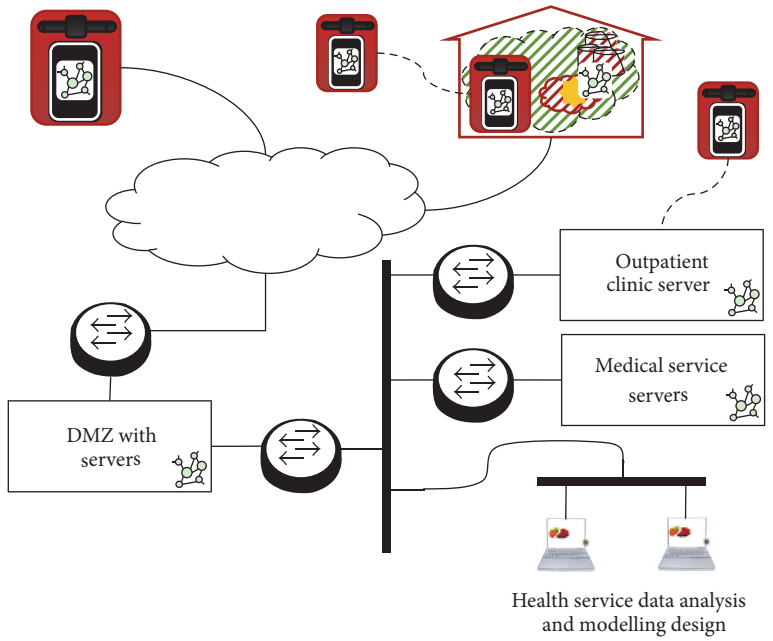


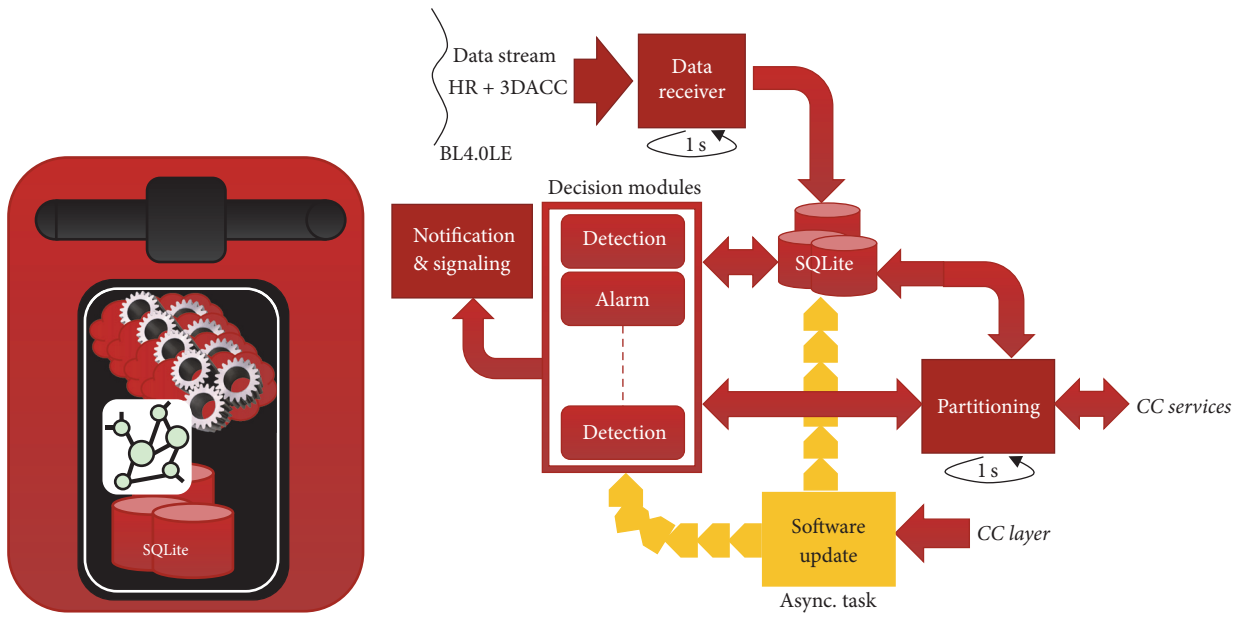
Figure 1: The architecture proposed in this study. Every single computing element has access to the ontology, the patient's tasks planning, and allocation for the current scenario. The PK can work alone or using the available CC services. Whenever possible, the configuration that leads to longer battery life will be chosen. CC services on the health service or in the specific medical unit service perform the calculations and data storage. Data visualization and report generation are typical in the data exploitation. Moreover, the medical unit should be able to perform data analysis and modelling using intelligent interfaces that integrate the needed ontology information in order to distribute the tasks.

those services that are required for a suitable performance of the epilepsy seizure detection and patient monitoring, while those services and tasks devoted to extract new knowledge should be carried out on specific servers belonging to the corresponding medical unit.

This abstract definition of the architecture needs completion: the ontology development and fulfillment, the infrastructure, the mediator implementation, and so on; however, it cannot be successfully detailed in a single study due to its own complexity. The remainder of this study focuses on a prototype developed that includes a MCC solution and the PK because it is the minimum part required that allows us to evaluate some parameters of the system. For this prototype, neither the ontology nor the extending capabilities have been introduced. However, this data gathering prototype is needed in order to obtain real data from epileptic patients, allowing us to develop the remaining modules.

3.3. The MCC and Monitoring Unit. As mentioned before, the PK includes a WD and a Smartphone. The sampling frequency of the sensory system should depend on the physical measurements: accelerometers need sampling frequencies higher than 10 Hz [3]; HR needs smaller sampling frequencies. However, the majority of the commercial Smartwatches or Smartbands do not allow apps to sample data from the sensory system: they only allow access to calculated transformations or induced variables. The WD manufacturers offer their own SDK, which may or may not allow reading the raw data from the sensor; in the majority of the cases, the sampling frequency might not be fixed. The majority of the products give access to a website where the aggregate variables are available, for instance, the burnt calories, but they do not store the instantaneous data. Besides, the main part of the HR sensors allows downloading the data but not by streaming, disabling the online process of the signal. Further requisites for the WD include the use of low energy consumption networking, such as bluetooth 4.0 Low Energy, and a valid battery duty cycle of about one day long. To our knowledge, the single marketed solution that was valid for this type of applications was Pebble; unfortunately, the company was acquired and closed by a competitor. Recently, Samsung's Gear 2 devices got their market price decreased, allowing them to be considered as valid candidates; nevertheless, this fact has happened while writing this study, so they have not been tested yet. Therefore, currently available solutions are not suitable for the PK; thus, in this study we proposed an ad hoc solution including 3DACM and HR; interested readers can get more information concerning this WD in Section 4.1.

The structure of this current proposal for MCC is mainly based on the challenges described in [13] for the definition of MCCs. The costs subsection is inspired by that of cost analysis detailed in [12]. Figure 2 shows both the scheme of the PK and the MCC architecture. The architecture design decision needs further detail; the next subsection gives details concerning this MCC layer. From Section 3.3.2 to Section 3.3.4, several design parameters are analyzed: the data partitioning reliability, the platform's privacy, the network and the energy efficiency, and the machine learning issues, correspondingly.



(a) The patient's kit scheme

(b) The MCC architecture

Figure 2: (a) The PK scheme. (b) The MCC architecture. In this study Android Smartphones have been used; therefore the SQLite database management system is used. The different services refer to the capacity of the software (i) to receive data from the WD and to send data to the CC layer, (ii) to receive and send notifications and alarms, (iii) to process the received data from the WD, and so on. The 1-second loops refer to tasks that are scheduled periodically. The software update also induces changes in the database; for instance, a Decision Tree classifier can retrieve its rule from the database. The changes in the software refer, then, to the improvements in the algorithm implementation to enhance the energy efficiency.

3.3.1. The MCC Architecture. Several different approaches for MCC architectures have been described in the literature, from the most centralized approaches, focused on very light clients and a powerful cloud part, like in the Centralized Cloud [11, 68], to the hyperdistributed and high network synchronization requirements, like cloudlet [69]. For the purpose of this study, we claim that an ad hoc decentralized solution (Mobile Ad Hoc NETworks, MANET [70]) is the most suitable architecture.

The MANET architecture provides a high level of autonomy, with the local storage capacity and data processing. Moreover, requests to heavy and computational expensive cloud services are allowed as well. Besides, the cloudlet solution is not suitable because there is a single wearable which might be connected to the MCC layer and, mainly, the data from each patient are totally independent of one another; this latter fact suggests that there is no need for MCC synchronization among MCC layers, relaxing the computational requirements of this layer.

Conversely, our proposal manages the Smartphone as a service node, being responsible for storing the instantaneous data and its transformations, performing low-medium computation tasks, and so on. Therefore, this solution is highly based on that of MANET Mobicloud [70], allowing distributed and collaborative CC services to offer their solutions when the network is available [71, 72].

Figure 2(b) shows the proposed architecture in detail. Two main tasks, the Data Receiver and the Partitioning tasks, are continuously dispatched by means of a timer. These tasks are in charge of the communications with the outer layers of the architecture and with the computer decision models (Decision Modules). The Data Receiver task gathers the data received from the bracelet, using the SQLite database for storing the data. This task receives the block of the measurements sampled in the bracelet during one second as input; these blocks are securely stored in the SQLite database or data repository.

The Partitioning task is in charge of the offloading as well; its aim is to partition the data and request CC or MCC services with either raw data or processed data. Whenever there is unprocessed data, this task performs sliding windows on the data, requesting services for computing data transformations and for performing decision models based on data. According to the energy efficiency information, the task will request services to the MCC or to the CC, storing the intermediate data in the SQLite if needed. Besides, the Partitioning task should be divided into two parts: one is the windowing service and the other is just a job scheduling task, responsible for the CC/MCC services request as well. In parallel, an ontology of services should be developed and deployed into the SQLite database.

Furthermore, the software update task is responsible for deploying any update in both the models and their parameters and of the scheduling of new MCC services, new decision models or data processing; it is performed on demand of the CC layers. Coordination between this task and the standard app updating would eventually be needed.

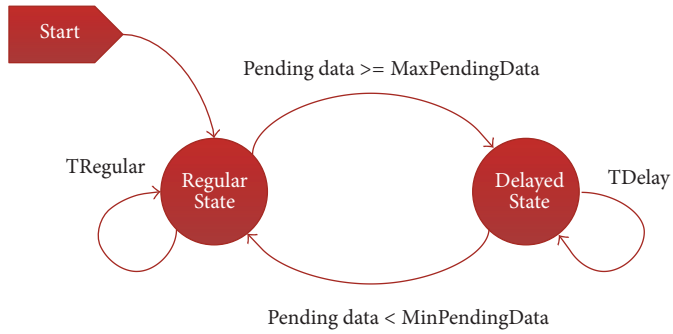


Figure 3: State diagram for the partitioning algorithm.

3.3.2. Reliability Issues in the Partitioning. This subsection focuses on three particular issues that IoT platforms must consider at least: the data delivery reliability, the data gathering service recovery, and the data delivery latency.

When talking about the two first issues, "reliability" refers to partitioning the data and requesting some services with the different partitions where the delivery of the data to the CC service must be guaranteed, while "recovery" deals with the healing of the crushed CC services and retrying the pending requests. Some of the solutions offered to cope with the reliability and the recovery of the services in the literature solve the problem at the expenses of the user experience [71], while other proposals like the Avatar system [73] suggest using a daemon that stays alive during network crashes keeping track of those partitions for which no service has been performed, retrying all the request just after the network connection is recovered. In this study, we propose an "Avatar"-like strategy, requesting the crashed services with the network recovery.

"Latency" is measured as the time elapsed since the request of a CC service on a partition until the answer or acknowledgment is received by the MCC. The performance of the system varies according to the data bunch size, ranging from 5 seconds to a few hours. Several factors have a great impact on the latency, for instance, the database's data storage mechanisms and the level of granularity of the services. Each of these factors needs further study in order to choose those that best fit the system performance in terms of real time and online monitoring. The experimentation in Section 4.2 analyzes the latency issue related to the MCC part.

MCC induces problems concerning the Quality of Service [17, 18]; the offloading techniques for partitioning are used to solve those problems related to the partitioning. Offloading varies from static partitioning of the data [74], aiming to reduce the battery consumption, to the dynamic partitioning [75, 76], where the partitioning is adjusted according to the network availability and the computing capacity of the mobile device. A hybrid partitioning is proposed using two defined data bunch sizes (NSRegular and NSDelayed). Each of them is used according to the network availability and its cost: if the WIFI networks are available then the NSRegular is used; otherwise the data is processed in the Smartphone and delivered in big bunches, of NSDelayed size, when the Wi-Fi network is present again. If the amount of undelivered data keeps growing up to MaxPendingData, then the data bunch size increases when Wi-Fi networks are available in order to deliver the data in the shortest period of time. Figure 3 shows the state diagram for the proposed partitioning.

The MCC computation and the data partitioning have a great impact on the energy expenditure and on the battery life: the smaller the computational effort in the Smartphone, the higher the network consumption. The reason for this compromise is that no computation in the MCC means delivering the instantaneous data to CC services; when using sampling frequencies higher than 1 Hz, the amount of data increases and so does the required communication bandwidth and complexity. Conversely, higher MCC computation induces a high reduction in the data that need CC delivery. Therefore, this research includes an energy efficiency experimentation to determine the best compromise in terms of how much computation can be offered by the MCC layer and how much are due to CC services.

The partitioning algorithm is described in Algorithm 1 where six global variables are set (lines (1)–(6)): DelayedState represents the state of working (initially it will be false), NSamples denotes the number of samples to send in each data bunch given the DelayedState, T is the delay time between two sent bunches, PendingData represents the number of samples to send, TempRepository is the local database, and finally WD represents the wearable device (Table 2 includes the description of the variables and values used in this algorithm). Two main tasks are scheduled (lines (7)-(8)): **DataReceiver** and **Partitioining** with a frequency of 1 and T seconds, respectively. The former deals with the communications with the bracelet, while the latter is devoted to the offloading and the data partitioning, including the partition type, RegularState or DelayedState. The DataReceiver running frequency is 1 second since it is the maximum frequency the WD can afford, while parameter T will be analyzed in Section 4.3.

The task **DataReceiver** (lines (10)–(19)) has to connect to WD through Bluetooth 4.0 with a discovering delay of TDiscovery seconds (line (16)). Once the connection is set, the data will be streamed from the WD with a frequency of one bunch per second. In parallel, the WD is sampling the HR and ACM sensors with a frequency of 16 Hz. In case the connection crashes, the lost samples will be skipped until the connection is recovered.

```
(1)  DelayedState ← false
(2)  NSamples ← NSRegular
(3)  T_DB ← TRegular
(4)  PendingData ← 0
(5)  WD ← Disconected
(6)  TempRepository ← []
(7)  Launch TimerTask DataReceiver() each 1 secs
(8)  Launch TimerTask Partitioining() each T_DB secs
(9)
(10) function DataReceiver
(11)     if WD is connected then
(12)         Sample ← ReadDataFrom(WD)
(13)         PendingData ← PendingData + 1
(14)         TempRepository ← [TempRepository Sample]
(15)     else
(16)         Sleep(TDiscovery)
(17)         WD ← BlueToothDiscovery()
(18)     end if
(19) end function
(20)
(21) function Partitioning
(22)     if PendingData > MaxPendingData and delayed == false then
(23)         NSamples ← NSDelayed
(24)         T_DB ← TDelay
(25)         DelayedState ← true
(26)     else
(27)         if PendingData < MinPendingData and delayed == true then
(28)             NSamples ← NSRegular
(29)             T_DB ← TRegular
(30)             DelayedState ← false
(31)         end if
(32)     end if
(33)     if CurrentBatteryLevel < BatteryThreshold then
(34)         Bunch ← TempRepository[1 ··· NSamples]
(35)     else
(36)         Bunch ← CalcTransformations(TempRepository[1 ··· NSamples])
(37)     end if
(38)     Launch sent ← SendData(Bunch)
(39)     if sent == true then
(40)         PendingData ← PendingData − NSamples
(41)         TempRepository ← Remove(TempRepository, NSamples)
(42)     end if
(43) end function
(44)
(45) function SendData(Bunch)
(46)     I ← NTry
(47)     sent ← false
(48)     while I < NTry and sent == false do
(49)         sent ← SendDataToCC(Bunch)
(50)         if sent == false then
(51)             Sleep(TRetry)
(52)         end if
(53)         return sent
(54)     end while
(55) end function
```

Algorithm 1: Partitioning and offloading algorithm.

Table 2: Variables used in the partitioning algorithm.

Variable	Description
TDiscovery	The timeout in the process of finding the bracelet Bluetooth device when the connection gets lost.
NSamples	The number of samples to send in each data bunch depending on the state of working: (i) NSRegular: the number of samples to send in each regular data bunch. A regular data bunch size is defined with the samples to deliver, measured in seconds. For instance, if ACM and HR are sampled at 16 bits and 16 Hz, one sampled data includes the (a) 16 bits $*3$ axis = 6 bytes from the ACM, (b) the 16 bits (2 bytes) due to the HR, and (c) 16 bits (2 bytes) reserved for future extension and the timestamp in 8 bytes; then the samples in a second are $16 * 18 = 288$ bytes. If NSRegular is set to the samples for 60 seconds, the regular data bunch size is $60 * 288$ bytes. (ii) NSDelayed: the number of samples to send in delayed state should be bigger than the NSRegular in order to go back to the Regular state.
T_{DB}	The time between two data bunches depending on the state of working: (i) TRegular: the time between two regular bunches, that is, the time between launching SendData tasks for the Regular state. (ii) TDelay: the time between two delayed bunches, that is, the time between launching SendData tasks for the Delayed state.
PendingData	Number of samples pending to send to the CC services: (i) MaxPendingData: maximum number of data samples that have not been completely acknowledged by the CC service, making the state change from Regular to Delayed. (ii) MinPendingData: threshold to change from Delayed to Regular states.
NTry	The number of retries of the SendData task.
TRetry	Time between two consecutive SendData task retries.
DelayedHours	Time spent in Regular state, given in hours; afterwards the process gets into the Delayed state.

The task **Partitioining** (lines (21)–(43)) updates the bunch size (NSamples) depending on the PendingData. When this variable surpasses the MaxPendingData threshold, the DelayedState is set to true and the related variables are updated. When PendingData is below MinPendingData, the DelayedState is set to false. Next, if CurrentBatteryLevel exceeds BatteryThreshold then the transformations of the first pending bunch in the TempRespository are calculated; see Section 4.1; otherwise, it is the raw data that is sent instead of its transformations (lines (33)–(37)). The bunch is sent to the CC services using the **SendData** function (line (38)) and in this case only the process succeeds, the PendingData is updated, and the bunch is removed from the TempRepository (lines (39)–(42)).

The auxiliary function **SendData** (lines (45)–(55)) will try to send the bunch to the CC services NTry times with a timeout of TRetry seconds (line (51)) between runs.

3.3.3. Security Issues within the Platform. Several ad hoc frameworks have been reported in the literature concerning interfaces between mobile devices and cloud services. To our knowledge, these solutions are not suitable for the specific problem focused on this research. Some of these frameworks, through BNS and IoT ad hoc solutions, are not secure [71, 77]; furthermore, its focus is a general platform, making it difficult to be extended for specific solutions [11]. Other solutions, like those in [70, 78], though being secure ad hoc frameworks, make use of weblets; therefore, the increased computational cost makes them fruitless [12]. Concerning the epilepsy specific frameworks, those reported in the literature are either very specific or closed solutions [7, 41], being isolated and storing all the patients' data in the Smartphone [32]. Thus, it was decided to develop our own framework to fit the specific needs of the problem, considering the extensibility and enhancement issues to improve the development.

The proposed architecture involves two kinds of vulnerable wireless connections: a bluetooth connection between the WD and the mobile phone and a wireless connection (Wi-Fi or data) between the mobile phone and the CC based on the REST protocol. The security and privacy issues on wearable communications are a challenging field of study [79, 80] since the computing capability and battery life of these devices are quite limited. In this sense, one of the most popular techniques for key generation/key agreement for wearable devices is based on physiological print such as the interpulse interval (IPI) [81]; however, our first version of WD does not include any kind of extra security issues since it will used in a secure and controlled environment for the first clinical tests. Besides, the wireless connection between the mobile phone and the CC is carried out using a RESTFul service together with the HTTP communication protocol [82], since we know the optimal solution would be an HTTPs connection.

3.3.4. Detection of Epileptic Seizures. Two different types of models are proposed for the detection of focal myoclonic epileptic seizures: on the one hand, Genetic Fuzzy Finite State Machines (GFFSM) applied to the epilepsy recognition [3] and, on the other hand, a feature extraction using a Distance-based Principal Component Analysis (DPCA) step followed by a K-Nearest Neighbor (KNN) classifier [65].

GFFSM defines two main fuzzy sets to describe the current state (seizure or normal) and a set of four fuzzy rules (IF-THEN fuzzy rules) whose output is the new fuzzy state. The ACM values are transformed to three new variables, Signal

Table 3: Bracelet and Smartphone specifications.

Bracelet		Smartphone	
Microcontroller	SoC BTLE nRF51822 (32-bit ARM Cortex M0 core) 256 kB RAM 32 kB FLASH	Platform and OS	32 bits, MediaTek MT6582, Android Lolipop 5.1
Accelerometry sensor	3-axial, 2G AD ADXL327BCPZ	CPU and core	1.3 Ghz Quad-Core ARM Cortex-A7
Pulse monitor	TI AFE4400RHAT	GPU	ARM Mali-400 MP2 500 Mhz
Pulse sensor	LED verde OSRAM SFH7050	RAM	1 GB
Battery	Ion-Lithium 425 mAh	Battery	Li-Po 2150 mAh
Antenna	PCB 2.4 GHz	Storage	8 GB

Magnitude Ratio (SMA), Amount of Movement (AoM), and Time between Peaks (TbP), which are the input variables, together with the current fuzzy State, to the fuzzy rule system. Provided that a good variable fuzzy partitioning algorithm is used [83], the GFFSM method produces highly generalized models to cope with a wide population [3].

Besides, the method published in [65] makes use of the ACM values, computing up to 23 different transformations, SMA, AoM, and TbP, among others, but transforms the domain to another one using DPCA. DPCA hybridizes Locally Linear Embedding (LLE) with Principal Component Analysis: the distance matrix is used to perform the PCA transformation instead of the covariance. In addition, the number of desired features in the transformed domain is given a priori, like in LLE; therefore high reduction in the dimensionality can be obtained. Applying DPCA to the dataset of all the known transformations for the ACM values, 23 features, and afterwards a K-NN classifier, with K set to 3, led to results similar to those obtained for GFFSM explained above.

These two options represent two totally different approaches: the former, with a reduced set of rules and states, stands for general models, valid for a wide population, that introduces simple computations; thus they can be easily performed in the MCC size. The main drawback of this method is the learning stage, which cannot be performed in MCC; however, some tuning and active learning issues can be considered in this context. On the other hand, DPCA + KNN introduces much more computation requirements in the detection service, but the learning and updating are absolutely affordable in terms of computational cost.

4. Numerical Results

This section deals with the studies concerning the latency and the energy efficiency described in the previous section (see Section 3). The aim of this experimentation is twofold: on the one hand, to determine the data bunch size for each of the states by means of analyzing the induced latency; on the other hand, the impact of MCC versus CC offloading in the duty cycle of the battery in a real context, so to choose the best energy efficiency balance.

The organization of this section is as follows. Firstly, the materials and methods used in this experimentation are detailed. Afterwards, the latency study is included in Section 4.2. Finally, the energy efficiency issues are presented in Section 4.3. In each of these subsections, a discussion on the findings is included.

4.1. Material and Methods. To carry on the experimentation, part of the detailed PK has been developed using an ad hoc bracelet; the core of the CC layer has also been implemented. The whole system is able to monitor the patient's behaviour, storing the data and performing transformations of the measured and sampled physical variables.

The ad hoc bracelet has been developed at the Instituto Tecnológico de Castilla y León [84], though the whole team participated in the design. It includes an Analog Devices ACM sensor and a Texas Instruments green light LED HR sensor; however, with the development of the wearable techniques, new sensors would eventually be introduced as well, extending the frontiers of the seizures and events to discover.

This bracelet delivers the sampled data to the linked Smartphone using bluetooth 4.0 Low Energy; the sampling frequency for the ACM has been set to 16 Hz, while the HR is based on reflected light and the measurements are gathered on demand with at least 10 seconds between consecutive estimations. The WD includes a 420 mA battery, which beats plenty of the solutions in the market and allows a day-long duty cycle The MCC is implemented using a centralized approach [11], which is depicted in Figure 4.

The Smartphone is a mid-class Android mobile device, with bluetooth 4.0 Low Energy capabilities. Obviously, the mobile is also provided with a Wi-Fi connection. The experimentation has been performed in a laboratory including two PKs, a Wi-Fi router, and a low-cost server located in a different laboratory, actually, in a another city, the two laboratories connected to the Internet. Table 3 includes the data specifications of the WD and the Smartphone, while Table 4 shows the specifications of the deployed server.

The MCC layer incorporates the Data Receiver and the Partitioning tasks, makes use of the SQLite database, and can accept requests to compute the ACM and HR defined transformations. Whenever the battery level is under a predefined

(a) WD frontal view

(b) HR sensor detail

Figure 4: (a) The ad hoc developed WD is shown, while (b) the green light LED on the bottom of the bracelet is shown.

Table 4: CC Server side specifications.

Server	HP ProLiant DL320 G4 2 GB RAM (Dual-Core Pentium D 900)
OS	Scientific Linux 6.3
HTTP server	Apache 2
REST services framework	SlimPHP 0.9
Database storage	QNAP TS412U, 4 × 1 TB HD, RAID1
Database server	MySQL 5.1.73, database on RAID1 with granted access through NFS

threshold, the raw data will be stored in the local database, requesting the further services when this condition is not held anymore. The CC layer can accept requests for storing and for processing data. Besides, a data unit sent from the bracelet to the Smartphone includes three axial components from ACM, the HR, an extra field, and the time stamp. All of them except the time stamp are 16 bits long; the time stamp is stored in 32 bits. The predefined transformations used in this experimentation include the SMA, the AoM, and the TbP, computed as reported in [3].

Two possible scenarios will be considered: the first, sending the raw data from the MCC to the CC; the second, performing the preprocessing in the MCC and sending the transformations to the CC. In both cases, JSON messages will be interchanged. Whenever raw data is interchanged between the MCC and the CC layers the JSON message is structured as shown in the upper part of Table 5, using the integer representation of the sampled physical variables.

```
(1) .....
(2) while I < NTry and sent == false do
(3)     Time_tic_start ← GetMilliseconds()
(4)     sent ← SendDataToCC(Bunch)
(5)     Time_tic_end ← GetMilliseconds()
(6)     Latency ← Time_tic_end − Time_tic_start
(7)     if sent == false then
(8)         Sleep(TRetry)
(9)     end ifreturn sent
(10) end while
(11) .....
```

Algorithm 2: Excerpt from partitioning main algorithm with the latency calculation specification.

A data bunch including samples for periods of 30 seconds long, therefore, will be of the size of 51,360 bytes.

Whenever the data sent from the MCC layer to the CC layer includes the transformations, the JSON format is as shown in the bottom part of Table 5. In this case, if a window size of 2 seconds is used, the data bunch would include 2755 bytes.

Finally, two experiments have been carried out. The former is related to analyzing the latency in the two defined states, RegularState and DelayedState, for different data bunch sizes when sending raw data from the MCC to the CC, with the aim of finding the best balance. The latency is measured during the data transmission as stated in Algorithm 2, which is an excerpt of the algorithm shown in partitioning but extended to include the measurement of the latency. Different sizes have been tested: 6.80 KB (samples for 5 seconds, a total of 80 raw samples), 13.59 KB (10 s, 160 raw samples), 20.39 KB (15 s, 240 raw samples), 40.78 KB (30 s, 480 raw samples), 81.56 K (1 min, 960 raw samples), 407.81 KB (5 min, 4800 raw samples), 2446.88 KB (30 min, 28800 raw samples), and 4893.75 KB (60 min, 57600 raw samples). Ten repetitions of the test have been run to obtain the statistical performance.

Once the best data bunch size is found, the second experiment aims to evaluate the best performance between MCC preprocessing and CC preprocessing. In the former case, the MCC layer computes the transformations and sends them in data bunches, while in the latter raw data is sent. In both cases, the same data bunch size is used: the one found optimum in the first experiment. The PK performs its normal operation from full charge to total discharge. Two series of runs have been run: one sending raw data to the CC and the other computing the transformations at MCC level and sending these transformations. Again, ten runs for each series have been performed. The next two subsections show the discussion on the obtained results for each experiment.

4.2. Discussion on the Results for the Latency Test. Results are shown in Table 6 and in Figures 5 and 6. Table 6 shows the obtained times of latency in milliseconds for each of the 10 runs of the experiment and for each data bunch size. Moreover, the mean and the standard deviation are also

Table 5: JSON interchanged messages. The upper message has been used when sending raw data, while the bottom one has been used when sending the transformations.

```
[
{"IDPatient":"1","Time":1484912812,"HR":"0","RESERVE":"0","ACCX":"399","ACCY":"413","ACCZ":"456"},
{"IDPatient":"1","Time":1484912812,"HR":"0","RESERVE":"0","ACCX":"399","ACCY":"413","ACCZ":"456"},
{"IDPatient":"1","Time":1484912812,"HR":"0","RESERVE":"0","ACCX":"399","ACCY":"413","ACCZ":"456"},
..... ]
```

```
[
{"IDPatient":"1","Time":1484912812,"HR":"98","Oxigen":"0","SMA":"1.98","AoM":"3.23","TbP":"4.56"}
]
```

Table 6: Latency obtained results, measured in milliseconds, for each data bunch size and iteration. The four bottom rows are the main statistics, the mean and the standard deviation, the time between data bunches (T_{DB}), and the ratio latency time versus T_{DB}.

Number of samples	80	160	240	480	960	4800	28800	57600
KB	6.80	13.59	20.39	40.78	81.56	407.81	2446.88	4893.75
Iter								
1	191	318	456	875	2082	8994	53234	105975
2	185	325	452	1116	2043	9973	52927	104775
3	155	305	467	1168	2057	9055	52594	110677
4	180	327	468	900	1967	9145	52296	105161
5	186	300	458	890	1844	9537	52585	106078
6	152	305	474	907	1965	9538	52619	105476
7	177	337	504	1154	1843	8916	52206	105004
8	182	565	464	1145	1956	8874	52413	104416
9	165	322	600	885	1825	8874	53141	104864
10	177	325	458	907	2083	8862	52312	104522
Mean	175,00	342,9	480,10	994,70	1966,50	9176,80	52632,7	105694,80
Std	13,28	78,90	44,61	130,97	100,75	379,17	340,15	1838,02
Latency/T_{DB}	0.035	0.034	0.032	0.033	0.033	0.031	0.029	0.029
T_{DB}	5	10	15	30	60	300	1800	3600

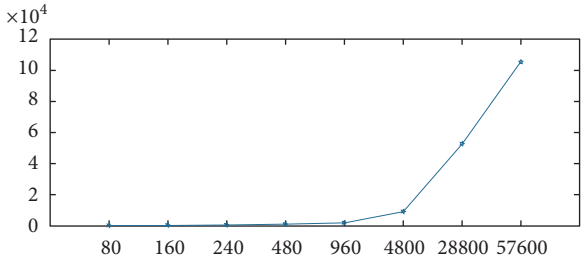


Figure 5: The exponential relationship between the data bunch size, x-axis, in KB, and the mean latency time, y-axis, in milliseconds, for the CC processing mode.

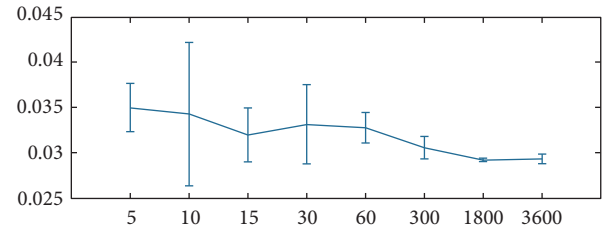


Figure 6: The relationship between the data bunch size and the ratio latency time versus the time between data bunch generation for the CC preprocessing mode. Some smaller sizes exhibit a wide spread; the higher the ratio, the narrower the variation in the performance.

calculated and shown for each size over the 10 runs. Finally, the time between consecutive data bunches (T_{DB}) and the ratio of the latency time versus T_{DB} are shown.

The relationship between the mean of the latency time and the data bunch size is shown in Figure 5, where it clearly shows two sizes as candidates: 20.39 KB, equivalent to 15 s gathering 240 raw samples, and 40.78 KB, equivalent to 30 s gathering 480 raw samples, represent the best compromise between the communication acts and the response time. Smaller sizes might introduce a shorter response but the communication acts will have a high impact on the battery life. Furthermore, the mean latency times, 480.10 ms and 994.70 ms, correspondingly, are still negligible compared to the time between consecutive bunches, 15 s and 30 s, allowing the recovery of connection lost events easily. More specifically, up to 14 retries can be securely done with a safe waiting period of 0.5 s (1 s for the second option) without overlapping, which can be considered enough in a real context.

Table 7: Performance facts when comparing MCC with CC services for the two analyzed data bunch sizes, 20.39 KB and 40.78 KB. For the 20.39 KB size, only the MCC has been performed as detailed in the text.

Parameter	20.39 KB	40.78 KB	
	MCC	MCC	CC
Battery life time (h)	28.998	32.232	31.600
MB of transferred data	6.625	8.702	159.223
Latency time (h)	4.491	4.555	51.214
Computational time (h)	1.041	1.395	0.000
Latency + computation (h)	5.532	5.950	51.214

These findings are also remarked in Figure 6. This figure shows the ratio latency time versus T_{DB} for all the iteration on each different data bunch size. For the smaller sizes, a high ratio variability is observed, which means that the reliability of the data delivery is in compromise. Higher values induce reduced risks, but at a cost of real time response. Considering the previous results and the ratio in this latter table, the data bunch size of 20.39 KB represents the best compromise.

4.3. Battery Duty Cycle Test. This test proposes putting the PK into deployment in the two main cases: with MCC computation and with all the computations carried out in the CC layer. In the former case, the communication acts deliver the transformation of the raw data, reducing the amount of data to send; the latter case includes delivering the raw data, reducing the computation at the expenses of increasing the amount of data to deliver. Two data bunch sizes are analyzed and used in this experimentation, those that found the best solutions in the previous subsection: 20.39 KB and 40.78 KB. We performed the CC and MCC calculations for 40.78 KB first; according to the obtained results, only the MCC calculations have been considered for the 20.39 KB case.

The results for these performance tests are included in Table 7 and Figures 7 and 8. Clearly, the main contribution to the battery consumption is the communication acts: the higher the computation in the Smartphone, the better. This conclusion only holds for those components that do not include human-machine interface, like the touchscreen. Besides, further analysis is required in order to evaluate the complexity of the calculations that do not introduce extra battery consumption. Finally, according to the evolution of the battery duty cycle, the most interesting data bunch size is 40.78 KB, as long as it provides the longest battery discharge time, while performing similarly concerning the latency. However, the 20.39 KB size can be considered valid as well, as no main differences have been found.

4.4. Deployment Cost of the Platform. The whole approach is based on relatively low-cost elements: a mid-range Smartphone plus a WD, the PK, worth less than 500 euros to the

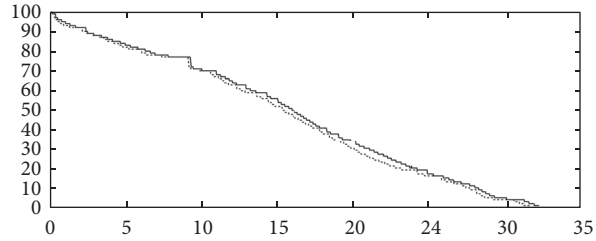


Figure 7: Comparison of the Smartphone's battery discharge when using the data bunch size of 20.39 KB for the two options: MCC preprocessing versus CC preprocessing. The continuous line stands for the MCC preprocessing mode, while the dotted line stands for the CC preprocessing mode. On the y-axis, the battery charge level is in percentage. The x-axis shows the time in hours.

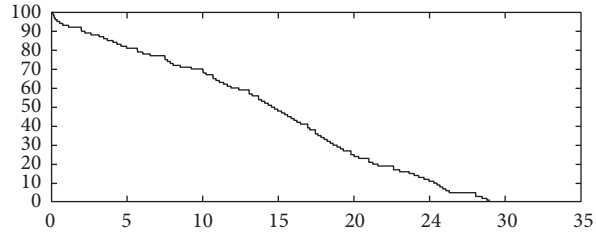


Figure 8: Comparison of the Smartphone's battery discharges when using the data bunch size of 40.78 KB for the MCC preprocessing. On the y-axis, the battery charge level is in percentage. The x-axis shows the time in hours.

health system and even less if we consider that the majority of the population already owns a valid Smartphone. The deployment of the CC services can be done either externally or in their own data center; the balance between MCC and CC services might help in reducing this deployment cost. Besides, the cost of Smartphones with the minimum requirements is normally 130 euros at the most. On the other hand, the wristband used in this study is worth 300 euros, but this price can be highly reduced; commercial Smartwatches integrating the HR and ACM are available from 150 euros.

It is worth mentioning that one of the main concerns with the wristband selection is the battery life: this parameter must be higher than 24 hours. A smaller battery life could lead to problems like a high number of charging cycles or too short a working time between charges. In any case, there is always the possibility of the battery running out, but this risk gets higher with reduced battery life periods. And this is something that can make the whole platform useless, increasing its opportunity cost.

Moreover, the integration of federated CC servers in outpatient clinics, even in family homes, with local servers under 500 euros nowadays, will highly decrease the amount of power required by the health system and outperforming both the robustness and the real time response. Not related to IoT but to eHealth, such a distributed solution has been successfully used in GNU Health [85]. This cost analysis has been barely performed; and therefore, the main part of the solutions is either expensive or uncomfortable, or even

both at the same time. Each design decision has been taken considering the cost analysis of the candidate solutions.

5. Conclusions

This study analyzes the solutions in the literature describing solutions for epilepsy tonic-clonic seizure detection and monitoring. The majority of the approaches lack several remarkable factors: developing ergonomic approaches, supporting everyday life, providing economical affordable solutions, introducing storage of the sampled data and providing intelligent CC services, introducing real time response, or considering multiple services on the MCC size. This study addresses the design of an IoT platform for the epilepsy seizure detection and monitoring considering each of these factors.

The solution is based on a WD to be located on a wrist connected to a Smartphone, which in turns implements MCC services and has access to CC services as well. The global goal is detecting the seizures, storing information from the sensory system, generating alarms and notifications, performing machine learning techniques on the data to learn the best models to detect or to visualize the data, sharing data, and providing processed information to the medical staff, among others. Special attention has been paid to the MCC module, where some design decisions are discussed, leading to the experimentation stage.

The experimentation stage implemented part of the MCC and CC modules, developing an ad hoc solution for the WD. The experimentation has been focused on determining the best data bunch size and on drawing conclusions concerning the criteria to choose when performing computation on the MCC versus requesting services on raw data to the CC layer. The experimentation results show two possible data bunch sizes (20.39 and 40.78 KB) as the most suitable ones. Furthermore, the second stage of the experimentation suggests that plenty of computation can be delivered on the Smartphones, reducing the amount of networking. Furthermore, special care should be taken to reduce the power consumption due to some mobile components, such as touchscreens. This research is only in its early stages, and in the near future we expect to complete the design, considering the integration of this framework into publicly available open software health platforms, such as GNU Health.

Abbreviations

ACM: Triaxial accelerometer
AoM: Amount Of Movement
BSN: Body Sensor Networks
CC: Cloud Computing
DPCA: Distance-based Principal Component Analysis
EEG: Electroencephalogram
ECG: Electrocardiogram
GFFSM: Genetic Fuzzy Finite State Machine
HR: Heart rate
KNN: K-Nearest Neighbor
MANET: Mobile Ad Hoc NETworks
MCC: Mobile Cloud Computing
PK: Patient's kit
SMA: Signal Magnitude Ratio
TbP: Time between Peaks
T_{DB}: Time between consecutive data bunches
WD: Wearable device.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research has been funded by the Spanish Ministry of Science and Innovation, under Project MINECO-TIN2014-56967-R, and Junta de Castilla y León, under Project BIO/BU01/15.

References

[1] R. S. Fisher, W. Van Emde Boas, W. Blume et al., "Epileptic seizures and epilepsy: definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, 2005.

[2] J. Engel Jr., "A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ILAE task force on classification and terminology," *Epilepsia*, vol. 42, no. 6, pp. 796–803, 2001.

[3] J. R. Villar, P. Vergara, M. Menéndez, E. De La Cal, V. M. González, and J. Sedano, "Generalized models for the classification of abnormal movements in daily life and its applicability to epilepsy convulsion recognition," *International Journal of Neural Systems*, vol. 26, no. 6, 2016.

[4] E. B. Petersen, J. Duun-Henriksen, A. Mazzaretto, T. W. Kjar, C. E. Thomsen, and H. B. D. Sorensen, "Generic single-channel detection of absence seizures," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS '11)*, pp. 4820–4823, Boston, Mass, USA, September 2011.

[5] C. Pradhan, S. Sinha, K. Thennarasu, and T. Jagadisha, "Quantitative analysis of heart rate variability in patients with absence epilepsy," *Neurology India*, vol. 59, no. 1, pp. 25–29, 2011.

[6] S. Beniczky, T. Polster, T. W. Kjaer, and H. Hjalgrim, "Detection of generalized tonic-clonic seizures by a wireless wrist accelerometer: a prospective, multicenter study," *Epilepsia*, vol. 54, no. 4, pp. e58–e61, 2013.

[7] D. Callegari, E. Conte, T. Ferreto et al., "EpiCare - a home care platform based on mobile cloud computing to assist epilepsy diagnosis," in *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare, MOBIHEALTH 2014*, pp. 148–151, November 2014.

[8] P. Lokhande and T. Mote, "Epilepsy monitoring and analysis system using android platform," *International Journal of Sciences and Research*, vol. 5, no. 7, pp. 60–64, 2016.

[9] P. Meghana and V. V. Sravani, "Mobile application based detection of seizures using inertial measurement units and emergency care," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 3, no. 11, pp. 6388–6391, 2015.

[10] S. Ramgopal, S. Thome-Souza, M. Jackson et al., "Seizure detection, seizure prediction, and closed-loop warning systems in epilepsy," *Epilepsy and Behavior*, vol. 37, pp. 291–307, 2014.

[11] S. Deshmukh and R. Shah, "Computation offloading frameworks in mobile cloud computing: a survey," in *Proceedings of the 2016 IEEE International Conference on Current Trends in Advanced Computing, ICCTAC 2016*, pp. 1–5, March 2016.

[12] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile cloud computing: a survey," *Future Generation Computer Systems*, vol. 29, no. 1, pp. 84–106, 2013.

[13] M. Othman, S. A. Madani, S. U. Khan et al., "A survey of mobile cloud computing application models," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 393–413, 2014.

[14] M. M. Hassan, H. S. Albakr, and H. Al-Dossari, "A cloud-assisted Internet of Things framework for pervasive healthcare in smart city environment," in *Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities, EMASC 2014*, pp. 9–13, ACM.

[15] G. Fortino, M. Pathan, and G. di Fatta, "BodyCloud: integration of cloud computing and body sensor networks," in *Proceedings of the 4th IEEE International Conference on Cloud Computing Technology and Science (CloudCom '12)*, pp. 851–856, IEEE, December 2012.

[16] A. Khelil, F. K. Shaikh, A. A. Sheikh, E. Felemban, and H. Bojan, "DigiAID: A wearable health platform for automated self-tagging in emergency cases," in *Proceedings of the 4th International Conference on Wireless Mobile Communication and Healthcare, MOBIHEALTH 2014*, pp. 296–299, IEEE, November 2014.

[17] H. T. Dinh, C. Lee, D. Niyato, and P. Wang, "A survey of mobile cloud computing: Architecture, applications, and approaches," *Wireless Communications and Mobile Computing*, vol. 13, no. 18, pp. 1587–1611, 2013.

[18] M. R. Rahimi, J. Ren, C. H. Liu, A. V. Vasilakos, and N. Venkatasubramanian, "Mobile cloud computing: a survey, state of art and future directions," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 133–143, 2014.

[19] C. Doukas and I. Maglogiannis, "Managing wearable sensor data through Cloud Computing," in *Proceedings of the 2011 3rd IEEE International Conference on Cloud Computing Technology and Science, CloudCom 2011*, pp. 440–445, IEEE, December 2011.

[20] C. Doukas and I. Maglogiannis, "Bringing IoT and cloud computing towards pervasive healthcare," in *Proceedings of the 6th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2012*, pp. 922–926, IEEE, July 2012.

[21] A. Forkan, I. Khalil, and Z. Tari, "CoCaMAAL: a cloud-oriented context-aware middleware in ambient assisted living," *Future Generation Computer Systems*, vol. 35, pp. 114–127, 2014.

[22] M. Chen, Y. Ma, S. Ullah, W. Cai, and E. Song, "ROCHAS: Robotics and cloud-assisted healthcare system for empty nester," in *Proceedings of the 8th International Conference on Body Area Networks, BODYNETS 2013*, pp. 217–220, October 2013.

[23] C. Doukas, T. Pliakas, and I. Maglogiannis, "Mobile healthcare information management utilizing cloud computing and Android OS," in *Proceedings of the 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '10)*, pp. 1037–1040, IEEE, September 2010.

[24] X. Huang and P. M. Matricardi, "Allergy and asthma care in the mobile phone era," *Clinical Reviews in Allergy and Immunology*, pp. 1–13, 2016.

[25] X. Liu, R. Wang, D. Zhou, and Z. Hong, "Smartphone applications for seizure care and management in children and adolescents with epilepsy: Feasibility and acceptability assessment among caregivers in China," *Epilepsy Research*, vol. 127, pp. 1–5, 2016.

[26] M. Chen, Y. Zhang, Y. Li, M. M. Hassan, and A. Alamri, "AIWAC: affective interaction through wearable computing and cloud technology," *IEEE Wireless Communications*, vol. 22, no. 1, pp. 20–27, 2015.

[27] H. Alemdar and C. Ersoy, "Wireless sensor networks for healthcare: a survey," *Computer Networks*, vol. 54, no. 15, pp. 2688–2710, 2010.

[28] 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.

[29] B. Direito, C. Teixeira, B. Ribeiro, M. Castelo-Branco, F. Sales, and A. Dourado, "Modeling epileptic brain states using EEG spectral analysis and topographic mapping," *Journal of Neuroscience Methods*, vol. 210, no. 2, pp. 220–229, 2012.

[30] S. Xie and S. Krishnan, "Wavelet-based sparse functional linear model with applications to EEGs seizure detection and epilepsy diagnosis," *Medical and Biological Engineering and Computing*, vol. 51, no. 1-2, pp. 49–60, 2013.

[31] S. Sareen, S. K. Sood, and S. K. Gupta, "An automatic prediction of epileptic seizures using cloud computing and wireless sensor networks," *Journal of Medical Systems*, vol. 40, no. 11, article 226, 2016.

[32] S. Sareen, S. K. Sood, and S. K. Gupta, "A cloud-based seizure alert system for epileptic patients that uses higher-order statistics," *Computing in Science and Engineering*, vol. 18, no. 5, pp. 56–67, 2016.

[33] A. Ulate-Campos, F. Coughlin, M. Gaínza-Lein, I. S. Fernández, P. L. Pearl, and T. Loddenkemper, "Automated seizure detection systems and their effectiveness for each type of seizure," *Seizure*, vol. 40, pp. 88–101, 2016.

[34] N. A. Marzuki, W. Husain, and A. M. Shahiri, "Myepipal, Mobile application for managing, monitoring and predicting epilepsy patient," in *Proceedings of the International Conference on Advances in Information and Communication Technology*, pp. 383–392, Springer, 2016.

[35] P. G. K. Prince, R. Hemamalini, and R. Immanuel Rajkumar, "LabVIEW based abnormal muscular movement and fall detection using MEMS accelerometer during the occurrence of seizure," *Indian Journal of Science and Technology*, vol. 7, no. 10, pp. 1625–1631, 2014.

[36] E. Schulc, I. Unterberger, S. Saboor et al., "Measurement and quantification of generalized tonic-clonic seizures in epilepsy patients by means of accelerometry-An explorative study," *Epilepsy Research*, vol. 95, no. 1-2, pp. 173–183, 2011.

[37] T. E. society, *Free Epilepsy Smart-Phone App*, 2013, https://www.epilepsysociety.org.uk/free-epilepsy-smartphone-app#.WC1jamSLR8c.

[38] V.-M. Teohari, C. Ungureanu, V. Bui, J. Arends, and R. M. Aarts, "Epilepsy seizure detection app for wearable technologies," in *Proceedings of the European Conference on Ambient Intelligence*, 2014.

[39] F. Masse, M. V. Bussel, A. Serteyn, J. Arends, and J. Penders, "Miniaturized wireless ECG monitor for real-time detection of epileptic seizures," *Transactions on Embedded Computing Systems*, vol. 12, no. 4, article 102, 2013.

[40] M. D. Kamat, M. P. S. Ganorkar, and M. R. Jain, "Child activity monitoring using sensors," *International Journal of Engineering and Techniques*, vol. 1, no. 3, pp. 129–133, 2015.

[41] G. Bajwa, R. Dantu, M. F. M. Issadeen, and R. M. Joseph, "Self-tracking via brain-mobile-cloud interface," in *Proceedings of the AAAI Spring Symposium Series*, 2013.

[42] K. G. Raam and R. Sasikala, "Detection and prediction of seizures using a wrist-based wearable platform," *International Journal of Control Theory and Applications*, vol. 9, no. 37, pp. 193–203, 2016.

[43] R. Sasikala, A. Asuntha, and S. Indirani, "Detection and prediction of seizures using a wrist-based wearable platform," *Journal of Chemical and Pharmaceutical Sciences*, vol. 9, no. 4, pp. 3208–3215, 2016.

[44] M. Velez, R. S. Fisher, V. Bartlett, and S. Le, "Tracking generalized tonic-clonic seizures with a wrist accelerometer linked to an online database," *Seizure*, vol. 39, pp. 13–18, 2016.

[45] H. Goelz, R. D. Jones, and P. J. Bones, "Wavelet analysis of transient biomedical signals and its application to detection of epileptiform activity in the EEG," *Clinical EEG and Neuroscience*, vol. 31, no. 4, pp. 181–191, 2000.

[46] M. Milošević, A. Van De Vel, B. Bonroy et al., "Detection of epileptic convulsions from accelerometry signals through machine learning approach," in *Proceedings of the 2014 24th IEEE International Workshop on Machine Learning for Signal Processing, MLSP 2014*, pp. 1–6, IEEE, September 2014.

[47] A. Van de Vel, K. Cuppens, B. Bonroy et al., "Long-term home monitoring of hypermotor seizures by patient-worn accelerometers," *Epilepsy and Behavior*, vol. 26, no. 1, pp. 118–125, 2013.

[48] A. Van de Vel, M. Milosevic, B. Bonroy et al., "Long-term accelerometry-triggered video monitoring and detection of tonic-clonic and clonic seizures in a home environment: pilot study," *Epilepsy and Behavior Case Reports*, vol. 5, pp. 66–71, 2016.

[49] Y. S. Ding and Z. Zilic, "ECG compression for mobile sensor platforms," in *Proceedings of the 13th Annual Body Sensor Networks Conference, BSN 2016*, pp. 99–104, June 2016.

[50] Neurosky. Eeg, neurosky mindband. 2009 http://store.neurosky.com/.

[51] Epdetect. A free mobile application to detect epileptic seizures with automated carer alert, 2009 http://www.epdetect.com/.

[52] R. S. Fisher, E. Bartfeld, and J. A. Cramer, "Use of an online epilepsy diary to characterize repetitive seizures," *Epilepsy and Behavior*, vol. 47, pp. 66–71, 2015.

[53] L. Irody, *Mobile Patient Diaries: Epidiary*, 2007, http://www.irody.com/mobile-patient-diaries/.

[54] Jiwok, *Epi & Me*, 2014, http://www.epiandme.com/.

[55] L. N. Ranganathan, S. Aadhimoolam Chinnadurai, B. Samivel, B. Kesavamurthy, and M. M. Mehndiratta, "Application of mobile phones in epilepsy care," *International Journal of Epilepsy*, vol. 2, no. 1, pp. 28–37, 2015.

[56] U. K. D. Roman and H. Lausen, *The Web Service Modeling Ontology Wsmo, Final Version*, 2005, Online; accessed 3-May-2017 http://www.wsmo.org/TR/d2/v1.1/.

[57] M. S. John Domingue and D. Roman, Web service modeling ontology (wsmo) - an ontology for semantic web services. 2005. Online; accessed 3-May-2017.

[58] D. Roman, U. Keller, H. Lausen et al., "Web service modeling ontology," *Applied ontology*, vol. 1, no. 1, pp. 77–106, 2005.

[59] M. Aazam, E. Huh, M. St-Hilaire, C. Lung, and I. Lambadaris, "Cloud of Things: Integration of IoT with Cloud Computing," in *Robots and Sensor Clouds*, vol. 36 of *Studies in Systems, Decision and Control*, pp. 77–94, Springer International Publishing, Cham, Switzerland, 2016.

[60] T. C. Arcadius, B. Gao, G. Y. Tian, and Y. Yan, "Structural health monitoring framework based on internet of things: a survey," *IEEE Internet of Things Journal*, vol. 4, no. 3, pp. 619–635, 2017.

[61] D. Gil, A. Ferrández, H. Mora-Mora, and J. Peral, "Internet of things: a review of surveys based on context aware intelligent services," *Sensors*, vol. 16, no. 7, article 1069, 2016.

[62] M. Hassanalieragh, A. Page, T. Soyata et al., "Health monitoring and management using internet-of-things (IoT) sensing with cloud-based processing: opportunities and challenges," in *Proceedings of the IEEE International Conference on Services Computing, SCC 2015*, pp. 285–292, IEEE, July 2015.

[63] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the internet of things: a survey," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 1, pp. 414–454, 2014.

[64] IBM. What is mobile cloud computing? 2017. Online; accessed 5-May-2017 https://www.ibm.com/cloud-computing/learn-more/what-is-mobile-cloud-com%20puting/.

[65] J. R. Villar, M. Menéndez, E. de la Cal, J. Sedano, and V. M. González, "Identification of abnormal movements with 3D accelerometer sensors for seizure recognition," *Journal of Applied Logic*, 2016.

[66] J. Alcalá-Fdez, A. Fernández, J. Luengo et al., "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2011.

[67] I. H. W. Eibe Frank and M. A. Hall, *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*, Morgan Kaufmann, 4th edition, 2016.

[68] F. Liu, P. Shu, H. Jin et al., "Gearing resource-poor mobile devices with powerful clouds: architectures, challenges, and applications," *IEEE Wireless Communications*, vol. 20, no. 3, pp. 14–21, 2013.

[69] M. Satyanarayanan, P. Bahl, R. Cáceres, and N. Davies, "The case for VM-based cloudlets in mobile computing," *IEEE Pervasive Computing*, vol. 8, no. 4, pp. 14–23, 2009.

[70] D. Huang, X. Zhang, M. Kang, and J. Luo, "MobiCloud: building secure cloud framework for mobile computing and communication," in *Proceedings of the 5th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2010*, pp. 27–34, IEEE, June 2010.

[71] R. Hasan, M. M. Hossain, and R. Khan, "Aura: An IoT based cloud infrastructure for localized mobile computation outsourcing," in *Proceedings of the 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2015*, pp. 183–188, April 2015.

[72] G. Huerta-Canepa and D. Lee, "A virtual cloud computing provider for mobile devices," in *Proceedings of the 1st ACM Workshop on Mobile Cloud Computing and Services: Social Networks and Beyond (MCS '10)*, ACM, June 2010.

[73] C. Borcea, X. Ding, N. Gehani, R. Curtmola, M. A. Khan, and H. Debnath, "Avatar: mobile distributed computing in the cloud," in *Proceedings of the 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2015*, pp. 151–156, April 2015.

[74] E. Cuervoy, A. Balasubramanian, D.-K. Cho et al., "MAUI: making smartphones last longer with code offload," in *Proceedings of the 8th Annual International Conference on Mobile Systems, Applications and Services (MobiSys '10)*, pp. 49–62, New York, NY, USA, June 2010.

[75] P. Angin, B. Bhargava, and Z. Jin, "A self-cloning agents based model for high-performance mobile-cloud computing," in *Proceedings of the 8th IEEE International Conference on Cloud Computing, CLOUD 2015*, pp. 301–308, July 2015.

[76] H. Wu and D. Huang, "MoSeC: Mobile-cloud service composition," in *Proceedings of the 3rd IEEE International Conference on Mobile Cloud Computing, Services, and Engineering, MobileCloud 2015*, pp. 177–182, IEEE, April 2015.

[77] G. Fortino, D. Parisi, V. Pirrone, and G. Di Fatta, "BodyCloud: a SaaS approach for community body sensor networks," *Future Generation Computer Systems*, vol. 35, pp. 62–79, 2014.

[78] E. E. Marinelli, "Hyrax: cloud computing on mobile devices using mapreduce," Technical Report, DTIC Document, 2009.

[79] X. Bellekens, A. Hamilton, P. Seeam, K. Nieradzinska, Q. Franssen, and A. Seeam, "Pervasive eHealth services a security and privacy risk awareness survey," in *Proceedings of the 2016 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, CyberSA 2016*, June 2016.

[80] A.-R. Sadeghi, C. Wachsmann, and M. Waidner, "Security and privacy challenges in industrial internet of things," in *Proceedings of the 52nd Annual Design Automation Conference, DAC '15*, pp. 54:1–54:6, ACM, New York, NY, USA, 2015.

[81] S. Wang, R. Bie, F. Zhao, N. Zhang, X. Cheng, and H.-A. Choi, "Security in wearable communications," *IEEE Network*, vol. 30, no. 5, pp. 61–67, 2016.

[82] J. H. Christensen, "Using RESTful web-services and cloud computing to create next generation mobile applications," in *Proceedings of the 24th ACM SIGPLAN Conference Companion on Object Oriented Programming Systems Languages and Applications*, pp. 627–633, ACM, October 2009.

[83] J. R. Villar, P. Vergara, E. D. La Cal, M. Menéndez, and J. Sedano, "Pre-clinical study on the detection of simulated epileptic seizures," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 24, no. Suppl. 2, pp. 33–46, 2016.

[84] Applied electronics and artificial intelligence department; Instituto Tecnológico de Castilla y León, 2017, http://www.itcl.es/.

[85] GNU health - freedom and equality in healthcare, GNU solidario, non-gubernamental organization, 2017, http://health.gnu.org/.

Hindawi
Journal of Sensors
Volume 2017, Article ID 3186482, 11 pages
https://doi.org/10.1155/2017/3186482

Research Article

VSMURF: A Novel Sliding Window Cleaning Algorithm for RFID Networks

He Xu,[1,2] Weiwei Shen,[1,2] Peng Li,[1,2] Daniele Sgandurra,[3] and Ruchuan Wang[1,2]

[1]*School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China*
[2]*Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks, Nanjing 210003, China*
[3]*Information Security Group, Royal Holloway, University of London, Surrey TW200EX, UK*

Correspondence should be addressed to He Xu; xuhe@njupt.edu.cn

Received 9 February 2017; Revised 24 May 2017; Accepted 22 June 2017; Published 27 July 2017

Academic Editor: Nashwa El-Bendary

Radio Frequency Identification (RFID) is one of the key technologies of the Internet of Things (IoT) and is used in many areas, such as mobile payments, public transportation, smart lock, and environment protection. However, the performance of RFID equipment can be easily affected by the surrounding environment, such as electronic productions and metal appliances. These can impose an impact on the RF signal, which makes the collection of RFID data unreliable. Usually, the unreliability of RFID source data includes three aspects: false negatives, false positives, and dirty data. False negatives are the key problem, as the probability of false positives and dirty data occurrence is relatively small. This paper proposes a novel sliding window cleaning algorithm called *VSMURF*, which is based on the traditional SMURF algorithm which combines the dynamic change of tags and the value analysis of confidence. Experimental results show that VSMURF algorithm performs better in most conditions and when the tag's speed is low or high. In particular, if the velocity parameter is set to 2 m/epoch, our proposed VSMURF algorithm performs better than SMURF. The results also show that VSMURF algorithm has better performance than other algorithms in solving the problem of false negatives for RFID networks.

1. Introduction

The Internet of Things (IoT) [1] is defined as the set of objects that can communicate over the Internet. Cloud computing is an Internet-based computing service that can provide task allocation [2] and secure data services [3] to IoT devices on demand [4]. This has led IoT to be a global technology used in many fields, such as ubiquitous cities [5]. With the development of these systems, data process technology has attracted many researchers' interest. As an example, to meet the secure requirements of ubiquitous cities, Shen et al. have proposed a sharing framework based on attribute-based cryptography to support dynamic operations for urban data [5]. Radio Frequency Identification (RFID) is considered one of the key technologies to realize IoT and is widely used in many areas, such as mobile payments, public transportation, health monitoring, environment protection, and smart city [6, 7]. More and more big data are generated by those types of RFID and Cloud-based systems, which make the spatial-temporal data management a current hot topic [8]. RFID uses radio frequency signals through wireless communications to achieve noncontact transmission of information and to perform automatic identification of objects attached with RFID tags. The advantage lies in the fact that the RFID tags and readers can perform identification without physical contact. RFID system can be divided into the following three components: readers, tags, and back-end computer system. The reader and tag communicate through an antenna. To this end, the reader firstly emits electronic signals through an antenna, and then the tag emits identification information of internal storage after receiving the signal. Secondly, the reader receives and identifies the information sent back from the tag via an antenna. Finally, the reader sends the identification results to the computer system [9]. The working principle of RFID system is shown in Figure 1.

In early RFID applications, the reader is directly connected to the application, and RFID data will be processed as logical data by the application. This data processing

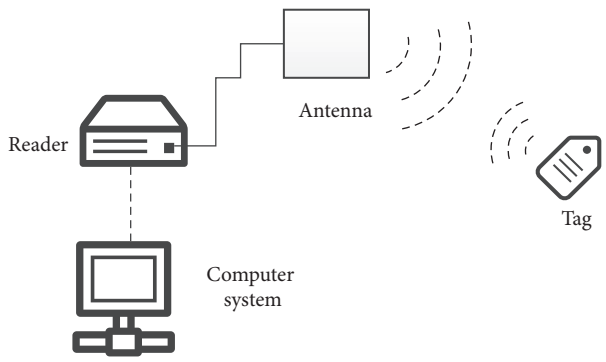


Figure 1: Working principle of RFID system.

approach meets the needs of earlier RFID systems, but the system's design is complex and the efficiency is low. There is an important problem with these legacy systems; in particular reusability is difficult. To reduce the complexity and meet the requirements of the rapid development of RFID technology, RFID middleware [10] systems were introduced. Middleware system works as a separate level which can perform some of the data processing and shield the reader hardware and the upper application system. In these scenarios, the program's scalability and applicability are greatly improved. RFID middleware can be used to clean, filter, and format the data collected by the readers and transfer the processed data to the back-end applications [11]. This paper mainly studies the data preprocessing method, which is also referred to as *data cleaning* method in RFID middleware system.

Traditional data cleaning technology [12] is mainly applied in three fields: (1) data warehouse (DW or DWH); (2) Data Discovery or Knowledge Discovery in Data (KDD) [13]; (3) Total Data Quality Management (TDQM) [14]. In these areas, data cleaning is an integral and essential part for data processing. However, traditional data cleaning technology is only suitable for relational database or data warehouse technology. The RFID data flow is different from the traditional relational database and the data stream generated in data warehouse. The interaction between the reader and the tag in the RFID system determines the following characteristics of RFID data [15, 16]: the source data has a simple structure that holds the characteristics of flow, batch, magnanimity, temporal, dynamic change, correlation, and unreliability. The way in which the data is generated by the RFID device determines that the source data is often seriously unreliable [17, 18]. The main reasons for unreliability of data in RFID system are as follows [19]:

(1) *False positives*: tags' data which should not be recognized by the reader for some reason (noise, electromagnetic interference, and the like) have been read.

(2) *False negatives*: tags' data that should be recognized by the reader have not been correctly read.

(3) *Dirty data*: reader detects the tag that exists in its reading range, but the data which is read by the reader contains errors.

False positive and dirty data are fortuity, and the probability of their occurrence is relatively small. Instead, the false negative phenomenon is more common and is the main reason of causing unreliability in RFID system. Therefore, in order to improve the quality of data and ensure that the upper application works effectively, RFID source data cleaning operation is needed. Previous researches aimed at RFID data cleaning technology have done extensive and in-depth study and proposed some classic data cleaning algorithms. This paper proposes a novel data cleaning algorithm (named as *VSMURF*), which is based on the traditional SMURF algorithm in RFID networks, which is aimed at reducing the number of false negatives.

The paper is organized as follows: Section 2 describes the related work and the SMURF algorithm. Section 3 presents the proposed VSMURF algorithm. Section 4 describes simulation and experimental results. Finally, Section 5 summarizes the paper.

2. Related Work and SMURF Algorithm

In this section we describe the related work on data cleaning technology and recap the SMURF algorithm.

2.1. Data Cleaning Technology. Data cleaning [20] refers to a process that can detect and correct the identifiable error in data files. Some classical data cleaning algorithms exist in the literature, which will be introduced in this section.

The data cleaning method based on sliding window is a typical and commonly used method. For example, Bai et al. proposed a method based on a fixed sliding window [21], where the window is fixed and moves forward over time. The cleaning process is shown in Figure 2. Here, the raw data is generated by reader.

Jeffery et al. proposed a scalable data cleaning framework [22] based on Extensible receptor Stream Processing (ESP), which is a declarative query processing tool that is easily pipelined to deploy the configuration to each recipient. The framework is capable of cleaning data from various features of different readers. ESP pipelined structure is processed into the following five consecutive programmable stages: Point, Smooth, Merge, Arbitrate, and Virtualize. Gonzalez et al. proposed a dynamic data cleaning method based on Dynamic Bayesian Networks (DBNs) [23], which dynamically adjusts the trustworthiness state (which is the probability of tag existence). Song of Liaoning University proposed a three-tier structure Kalman Filter-Based RFID Cleaning (KFBC) [24] to address the problem of false negative in RFID system. As shown in Figure 3, this is a Kalman filter update process, which consists of time update and measurement update, and it is an autoregressive process.

2.2. SMURF Algorithm. Here, we firstly recall the basic concepts of the SMURF algorithm [25], and then we describe the algorithm in detail.

Interrogation Cycle. It is an inquiry and answer-response process between a reader and a tag, which is the basic reader's protocol that tries to detect all tags by the reader.

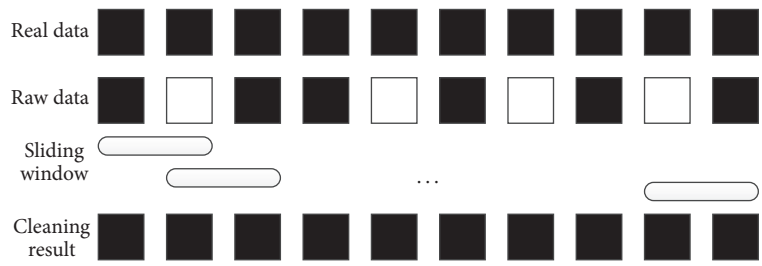


Figure 2: Data cleaning method based on fixed sliding window.

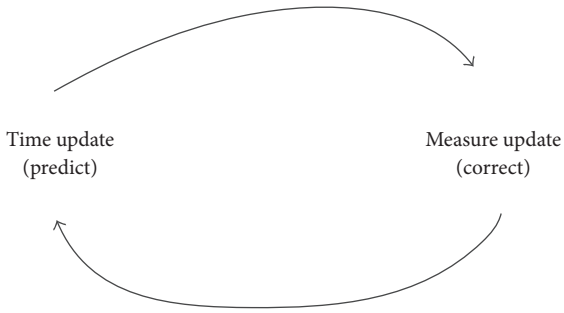


Figure 3: The updating process of Kalman filter.

Reading Cycle. It is a set of multiple interrogation cycles. It is also named as epoch, the value of which is from 0.2 seconds to 0.25 seconds [25]. In each epoch, the reader keeps the recording of how many and which tags are identified.

Read Rate $p_{i,t}$. It is the probability of tag i to be read by reader at t epoch, which is calculated as

$$p_{i,t} = \frac{\text{Responses}}{N_t}, \tag{1}$$

where N_t is the number of interrogation cycles at t epoch, and Responses is the number of responses of tag i at t epoch in the reader's reading range.

Sliding Window Size W_i. The value denotes w_i epochs in which tags can be identified by readers. It can be expressed by $W_i = (t - w_i, t]$.

S_i. It denotes the subset of tags observed by the reader during that epoch, which can be expressed by $S_i \subseteq W_i$. In the general case, it is assumed that tag i is seen in a subset of all the epochs in W_i.

The SMUR algorithm works as follows. In each epoch, the reader detects all tags in its reading range and stores the information of tags into the tag list, including *TagID*, *Responses*, and *Timestamp*, as shown in an example in Table 1. The list can be represented by a triple (TagID, Responses, and Timestamp), and all those pieces of information are transmitted to the reader's client at regular intervals.

Each epoch is treated as an independent Bernoulli experiment. In detail, the probabilities of tag i appearing in the window size (W_i) at each epoch are all the same and can be calculated using the formula:

Table 1: List of tags.

TagID	Responses	Timestamp
8576 4432 1345 5343	9	11:05:09
8576 9533 2334 2846	1	11:05:09
8576 9864 3453 8642	7	11:05:10

$$p_i^{\text{avg}} = \left(\frac{1}{|S_i|}\right) \cdot \sum_{t \in S_i} p_{i,t}. \tag{2}$$

In addition, the following two formulas are the most important ones in the SMURF algorithm:

$$W_i \geq \left\lceil \frac{1}{p_i^{\text{avg}}} \ln\left(\frac{1}{\delta}\right) \right\rceil, \tag{3}$$

$$\left||S_i| - W_i \cdot p_i^{\text{avg}}\right| \geq 2 \cdot \sqrt{W_i \cdot p_i^{\text{avg}} (1 - p_i^{\text{avg}})}, \tag{4}$$

where δ denotes *confidence* in formula (3). Confidence method can be used to process and analyze various data, such as image reconstruction. In the literature [26], authors use the confidence method to remove the noise from image according to the confidence of the depth to compute the range map. Experimental results show that the confidence method obtains good effectiveness in data processing. In SMURF algorithm, confidence is used to meet the integrity requirements of the data and ensure that the tag is successfully read in the sliding window W_i according to formula (3). Formula (4) is obtained by the central limit theorem, which is the control condition for setting the current sliding window size and the conditions of the tag's dynamic change. A reasonable sliding window size needs to balance the integrity requirements with the tag dynamic changes. Determining whether to output data is decided according to the data in the window in SMURF algorithm. In addition, determining how to adjust the window size is done according to the requirements of integrity and dynamic detection. Therefore, SMURF is a type of single tag's data cleaning algorithm.

SMURF algorithm is composed of the following steps.

Step 1. The initial setting of W_i is 1. T is the tag's pool and W_i^* is calculated according to formula (3).

Step 2. If $W_i^* > W_i$, adjust the window size according to the following formula:

$$W_i \longleftarrow \max\left\{\min\left\{W_i + 2, W_i^*\right\}, 1\right\}. \quad (5)$$

Step 3. If $W_i^* \leq W_i$, the dynamic change state of tag is detected according to formula (4). If the tag's state is changed, the window size needs to be adjusted according to the following formula:

$$W_i \longleftarrow \max\left\{\min\left\{\frac{W_i}{2}, W_i^*\right\}, 1\right\}. \quad (6)$$

2.3. Limitations of SMURF Algorithm. In the RFID middleware system, the sliding window filtering method is the key technology to reduce false negatives. However, there are some shortcomings in SMURF algorithm: firstly, the size of sliding window is difficult to set in the case of dynamic movement of tags. If the window is set too large, it will generate false positives in the process of filling data; and if the window is set too small, it will not be able to completely fill the data which results in false negative. The selection of the optimal value of sliding window size directly affects the effectiveness of the algorithm. In addition, the effect of this cleaning algorithm is remarkable only when the RFID data stream is in an ideal condition, which means that the RFID tags move at a uniform speed. However, this situation does not happen in some conditions, for example, when the external environmental factors change (such as the tag moves in or out at the reader's detection range rapidly). In these cases, the performance of the cleaning effect is greatly reduced. In addition, SMURF algorithm's confidence is just based on empirical evaluation; that is, it is not the result of a specific analysis. If the tag quickly leaves the reader's detection range, the value δ has a great impact on the results [20].

Specifically, formula (3) shows how the sliding window size (W_i) is related to the two parameters p_i^{avg} and δ. In fact, W_i is inversely proportional to p_i^{avg}, which means that the performance of the SMURF algorithm decreases rapidly when the tag moves out or into the reader's detection range. Assuming a set of raw data streams (0.5, 0.6, 0.6, 0.6, 0.5, 0.8, 0.8, 0.0, 0.4, and 0.8), the initial window size (W_i) is set to 5 epochs, where each epoch contains 10 interrogation cycles, and δ is set to 0.05. Hence, after calculation, p_i^{avg} is 0.56. Because the sliding window size (W_i) is 5 epochs, the processing tag has been moved out from the reader's detection range in the third epoch of the second window. In the end, the tag is still in reader's detection range according to SMURF algorithm, which results in false positives in RFID network. Therefore, SMURF algorithm fails under this condition. For these reasons, we have designed a novel algorithm to take into account the limitations of SMURF, which is described in the next section.

3. VSMURF: A Novel Sliding Window Cleaning Algorithm for RFID Networks

Based upon the previous analysis, we propose a novel algorithm to address SMURF algorithm's limitations, which is called VSMURF. In the following, we firstly describe the dynamic detection mechanism of tags and analysis of confidence. Then, we detail our adaptive sliding window cleaning algorithm.

3.1. The Dynamic Detection Mechanism of Tags. Assume that the current window size of tag i is $w_i = (t - w_i, t]$, w_i is divided into two parts, where the first part is denoted as $w_{1i} = (t - w_i, t - w_i/2]$ and the second part is denoted as $w_{2i} = [t - w_i/2, t]$, and the binomial distribution samples are, respectively, denoted as $|S_{1i}|$ and $|S_{2i}|$.

The dynamic detection mechanism of a tag's movement is divided into three parts.

Mechanism 1. In order to determine when the tag leaves the reader's detection range, the algorithm introduces the sliding window w_{2i} ($|S_{2i}|$) based on formula (4). The Least Squares Method [27] is used to fit the slope of the curve to determine whether the tag is moving out. If the slope is negative, the tag is moving out from the detection range. In addition, our proposed algorithm reduces two epochs of sliding window size. The purpose of this is to reduce the occurrence of false negatives.

Mechanism 2. The sliding window will be reduced by half if the tag is moving out and cannot be detected in the sliding window w_{2i} ($|S_{2i}| = 0$).

Mechanism 3. The sliding window size is increased if w_i^*, which is calculated according to formula (3), is larger than the current window w_i ($w_i^* > w_i$) and the actual number of readings is greater than the expected number of readings ($|S_i| > w_i \cdot p_i^{\text{avg}}$).

The above three mechanisms are used to adaptively adjust the size of the tag's sliding window. The tag's dynamic detection still uses the central limit theorem (CLT) [28]. In the initial state, all tags which need to be detected are set to 1 epoch and the window size is adjusted to 3 epochs. The goal is to balance the efficiency of the proposal algorithm and reduce false positives. Similar to the SMURF algorithm, the proposed VSMURF algorithm outputs the results at the middle of window and slides once at each epoch.

3.2. Confidence δ Analysis. In the SMURF algorithm, the formula $(1 - p_i^{\text{avg}})^{w_i} < \delta$ is the integrity requirement to ensure that the RFID data stream is completely covered, but it does not cover the particular value of δ that should be selected. The SMURF algorithm only provides an empirical value. When the value of δ is less than 0.5, it has a little effect in the sliding window size [25]. In fact, the SMURF algorithm cleans the data significantly when the tag moves slowly and does not move in and out frequently in the reader's detection range. But when the tag moves quickly, the algorithm's error rate becomes high. The reason is that the sliding window size becomes larger when the confidence δ is smaller, which results in an increase in the number of false positive readings and a corresponding increase in reading error rate.

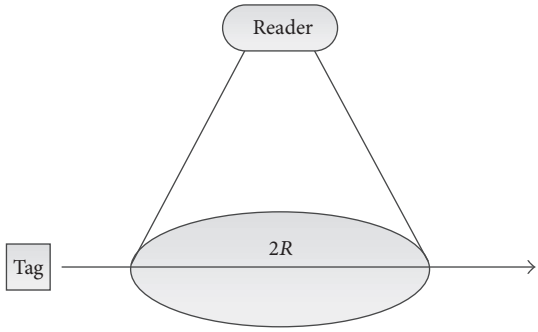


Figure 4: Reader detection range.

In order to solve the above problems, this paper takes into account the following factors that affect the cleaning results: the reader's detection range (R), reading frequency (f), the tag's speed (V), and confidence (δ). All these factors dynamically affect the sliding window size's adjustment. The maximum number of epochs is determined by $2R/(V/f) = 2Rf/V$ when the tag passes through the center of the circle (as shown in Figure 4).

When the tag is very close to the antenna of the reader reading range, the number of readings is $F = Rf/V$. To improve the efficiency of data cleaning, VSMURF algorithm dynamically adjusts the parameter δ ($\delta = p_i^{\text{avg}}/F$). Therefore, we obtain the following formula:

$$\left(1 - p_i^{\text{avg}}\right)^{W_i} < \delta = \frac{p_i^{\text{avg}}}{F} = \frac{p_i^{\text{avg}} V}{Rf}. \tag{7}$$

By applying the natural logarithm on both sides of the equation, we get

$$W_i \ln\left(1 - p_i^{\text{avg}}\right) < \ln\left(\frac{p_i^{\text{avg}} V}{Rf}\right). \tag{8}$$

This is because

$$\ln(1 - x) \le -x, \quad x \in (0, 1). \tag{9}$$

Finally, we obtain

$$W_i \ge \frac{\ln\left(Rf / p_i^{\text{avg}} V\right)}{p_i^{\text{avg}}}. \tag{10}$$

Thus, the integrity requirement is adjusted to formula (10). The maximum threshold υ of V can ensure that the error rate is less than 10% [20].

3.3. Adaptive Sliding Window Cleaning Algorithm VSMURF. VSMRUF is an improved single tag's sliding window cleaning algorithm based on SMURF algorithm and is based on the adjustment process of tag i at each epoch. The algorithm is composed of seven steps, which are described in detail in the following.

Step 1. Initialize the reader's detection range (R), reading frequency (f), confidence (δ), and threshold of tag's maximum speed and set the initial window to 1 epoch. When $V > \upsilon$, δ is increased appropriately ($\delta < 0.25$). Otherwise, δ is reduced appropriately.

Step 2. Detect whether reading cycle (epochs) ends, and if it ends, then the algorithm ends.

Step 3. Calculate the minimum value that satisfies the integrity requirement according to formula (3) and determine whether the tag is removed by using least squares.

Step 4. If the tag is moving out and $|S_{2i}| = 0$, the sliding window size is adjusted to the half according to Mechanism 2. Then adjust the window size according to formula (11) and enter the next epoch:

$$W_i \longleftarrow \max\left\{\min\left\{\frac{W_i}{2}, W_i^*\right\}, 3\right\}. \tag{11}$$

Step 5. Check whether the window is modified according to formula (4). If the window transforms, adjust the window size according to formula (12) and enter the next epoch:

$$W_i \longleftarrow \max\left\{\min\left\{W_i - 2, W_i^*\right\}, 3\right\}. \tag{12}$$

Step 6. If the current window size does not meet the integrity requirement ($W_i^* > W_i$) and the window does not convert ($|S_i| > W_i \cdot p_i^{\text{avg}}$), according to Mechanism 3 and to formula (13) increase the window size and then enter the next epoch:

$$W_i \longleftarrow \min\left\{W_i + 2, W_i^*\right\}. \tag{13}$$

Step 7. If the current window satisfies the integrity requirement ($W_i \ge W_i^*$) and the window transforms ($|S_i| \le W_i \cdot p_i^{\text{avg}}$), the window size will not change.

Figure 5 shows the data process flow of VSMURF algorithm.

The pseudocode of the VSMURF algorithm is given in Pseudocode 1.

4. Simulation and Experimental Results

4.1. Reader's Detection Model and Evaluation Mechanism. In our experiments, the detection range of the reader is divided into the major detection range and the minor detection range [29] (as shown in Figure 6).

The reader detection model has the following characteristics:

(1) The reader has a high detection probability and the reading rate is higher than 95% in the major detection range, which is near to the reader.

(2) The area that extends from the end of the major detection range to the end of the reader's detection range is called minor detection range. Here, the reading rate is linearly reduced to 0.

The model uses the following parameters to capture the behavior of readers under different conditions:

(i) *Detection range:* the distance between the reader and the boundary of the reader

(ii) Percentage of major detection range *(MajorPercentage)*: the major detection range of the reader accounting for the percentage of the entire detection range

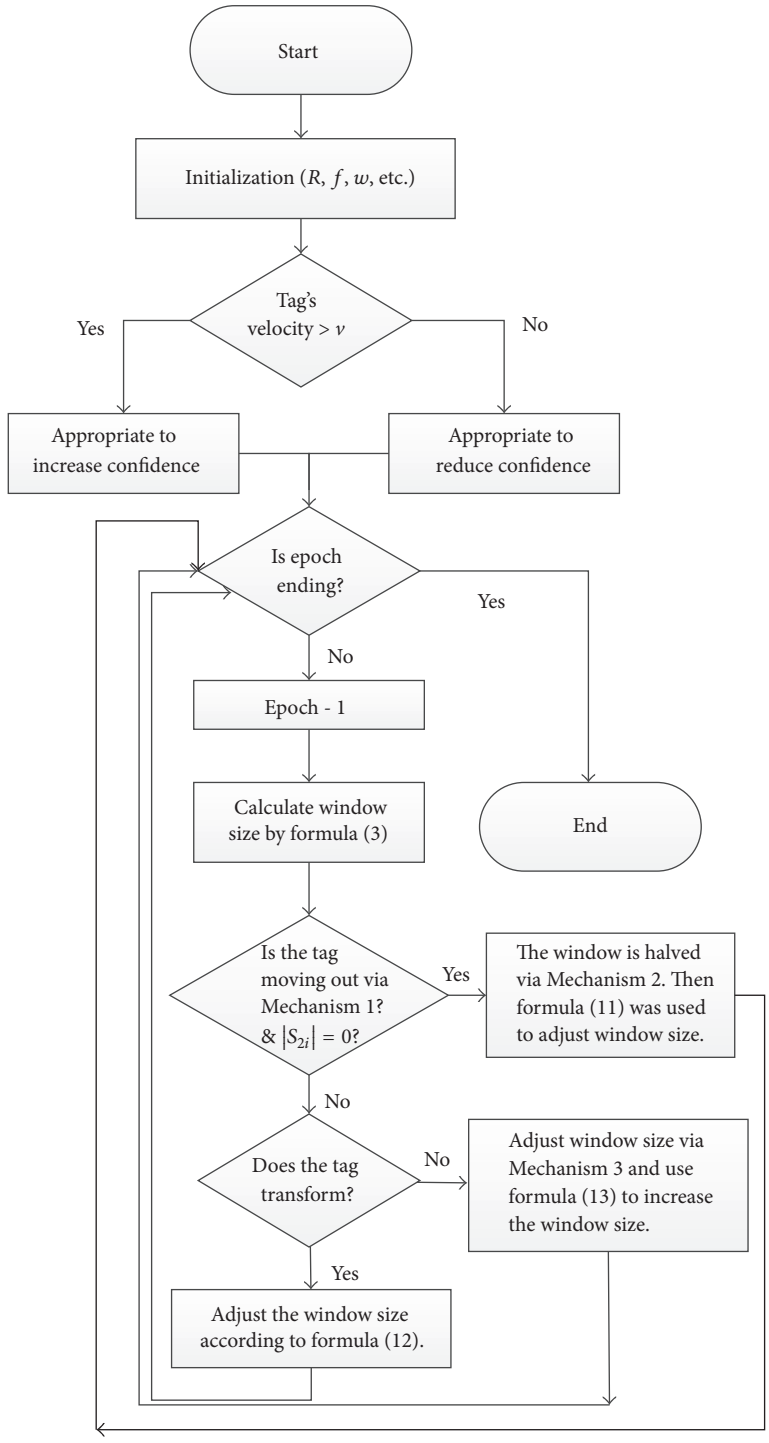


Figure 5: Data process flow of VSMURF algorithm.

(iii) Major reading rate *(MajorReadRate)*: reading rate of tags in the major detection range

(iv) *Error rate*: error data divided by the total number of data

In addition, evaluation mechanism of RFID data cleaning algorithm mainly uses the average error rate (AvgErrors), the average false positive readings (AvgPositives), and the average false negative readings (AvgNegatives). They are calculated as follows:

$$\text{AvgErrors} = \frac{\sum_{j}^{\text{NumEpochs}} \left(\text{FalsePositives}_j + \text{FalseNegatives}_j\right)}{\text{NumEpochs}},$$

```
Input: T = set of all observed tag IDs
       δ = required completeness confidence
       υ = the maximum threshold of tag's speed
       R = the radius of reader's detection range
       f = the frequency of reader's detection
Output: t = set of all present tag IDs
Initialize: ∀i ∈ T, W_i ← 1, δ = 0.1
While (getNextEpoch( )) do
  For (i in T) do
    processWindow(W_i) → p_{i,t}, p_i^avg, |S_i|, V
    if((V > υ) ∧ (V < 0.25))
      δ+ = 0.05
    Else
      δ− = 0.05
    End if
    If(tagExist(|S_i|))
      Output i
    End if
    W_i^* ← completeWindowSize(p_i^avg, δ)
      tagleaving ← moveDetection(p_{i,t}, S_i)
    If(tagleaving ∧ (|S_2i| = 0))
      W_i ← max{min{W_i/2, W_i^*}, 3}
    Else if(detectTransition(|S_i|, W_i, p_i^avg))
      W_i ← max{min{W_i − 2, W_i^*}, 3}
    Else if((W_i^* > W_i) ∧ (|S_i| > W_i · p_i^avg))
      W_i ← min{W_i + 2, W_i^*}
    Else
      W_i ← min{W_i, W_i^*}
    End if
  End for
End while
```

Pseudocode 1: Adaptive sliding window cleaning algorithm.

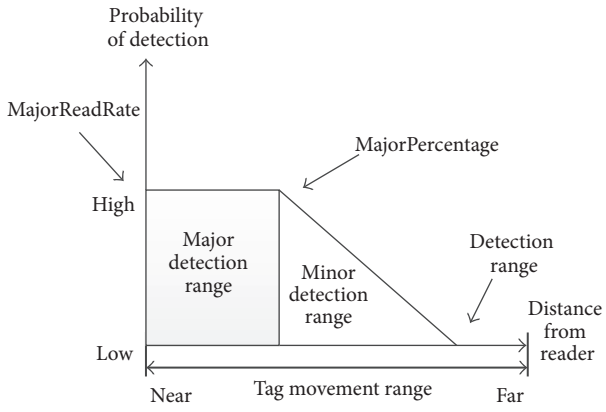


Figure 6: Reader detection model.

$$\text{AvgPositives} = \frac{\sum_{j}^{\text{NumEpochs}} \left(\text{FalsePositives}_j\right)}{\text{NumEpochs}},$$

$$\text{AvgNegatives} = \frac{\sum_{j}^{\text{NumEpochs}} \left(\text{FalseNegatives}_j\right)}{\text{NumEpochs}}, \tag{14}$$

where NumEpochs is the number of reading cycles; FalsePositives and FalseNegatives are the numbers of false positive readings and false negative readings, respectively. The evaluation mechanism includes two types of errors, which can effectively evaluate the performance of various algorithms.

The reading rate p is defined as in the following formula:

$$p(D) = \begin{cases} 1 - \dfrac{1 - p_h}{r_2} * D, & D \in (0, r_1), \\ \dfrac{p_l}{r_1 - r_2} * D - \dfrac{p_l * r_2}{r_1 - r_2}, & D \in (r_1, r_2), \end{cases} \tag{15}$$

where D is the distance between the tag and the reader, r_1 and r_2 are the reader's major and minor detection range, and p_h and p_l are the reading rates of the major and minor detection range, respectively.

4.2. Settings of Simulation. This section describes the experimental hardware and software environment used in all the simulations.

4.2.1. Hardware Environment. The hardware is composed of the following equipment:

(i) PC is HP Intel Core 2 Duo notebook. The speed of CPU is 2.33 GHz and the memory is 2 G.

(ii) RFID Reader is Speedway R420 reader.

(iii) Tag is EPC G2 UHF RFID tag.

Figure 7 shows the RFID reader, reader's antenna, and tags used in our experiments.

4.2.2. Software Environment. The software environment includes the following:

(i) Operating system: Windows 7

(ii) Programming language: Java

(iii) Programming environment: Java JDK 1.8.0_40; Eclipse IDE for Java Developers Luna Service Release 2 (4.4.2)

VSMURF algorithm belongs to the single tag cleaning algorithm; hence, in the experiments, only one reader is used and the number of tags is set to 25. The experimental model of VSMURF algorithm is shown in Figure 8. At some point, when the tag moves into the reader's detection range, p is calculated on the current coordinate point. The value of p is determined by the distance between the tag and the reader using formula (15). In addition, the system generates a random number n (the value of is n between 0 and 1), where n is used to determine whether the tag's information is generated. If $p \geq n$, the tag is read; if $p < n$, then the tag is false negative reading.

4.3. Results and Discussion. In this section we describe the experiments that we have performed.

(a) RFID reader

(b) Reader's antenna

(c) Tag

Figure 7: RFID reader (a), reader's antenna (b), and tag (c).

Table 2: Experiment 1 parameter setting.

Parameters	Value
Detection range	20 m
MajorPercentage	Variable (0, 1)
MajorReadRate	0.8
Number of tags	25
Speed	2 m/epoch
Reading cycles	1000

Experiment 1. 25 tags are selected, where each tag randomly changes its state, the MajorPercentage is changed from 0 to 1, and the reading cycles are set as 1000 epochs. We adopt and compare the following three algorithms: fixed window cleaning algorithm, SMURF cleaning algorithm, and VSMURF algorithm. The average error rate per epoch of above algorithms is compared. The experiment parameters are listed in Table 2.

The results are shown in Figure 9. In detail, when MajorPercentage is 0 (such as in a very noisy environment), the large window performs better, which shows about 4 errors per epoch. As MajorPercentage increases, the reliability of the original data and the accuracy of each algorithm are improved. The small window performs better when MajorPercentage reaches 1. For example, the major detection range covers the entire detection range of the reader. In this experiment, the VSMURF cleaning algorithm has a noticeable efficiency improvement compared with the SMURF algorithm and has the lowest error rate of the whole detection range at each epoch.

Table 3: The execution time of SMURF and VSMURF algorithm.

Algorithm	Execution time
SMURF	16.33 s
VSMURF	20.76 s

In order to simulate realistic conditions in terms of tag and reader behavior, we have implemented the SMURF and VSMURF algorithm using R420 RFID readers. In the experiments, we have collected 12 sets of real data. Due to the number of reading cycle, which is set to 1000, the time of data cycles is equal 1000. In the experiments, each set of data needs to record the execution time. The time result is shown in Table 3. From Table 3 and Figure 9, we can see that when the tag speed is large and the tag is moved out of the reader detection range frequently, the execution time of VSMURF is slightly higher than that of the SMURF algorithm because the VSMURF algorithm adds the confidence judgment. However, the average error rate of VSMURF is lower than that of SMURF algorithm.

In addition, the overhead of VSMURF algorithm with respect to traditional fixed window and SMURF algorithm is minimal because our VSMURF algorithm only utilizes simple mathematical operations.

Experiment 2. 25 tags are selected, and the tag's speed is changed randomly, the MajorPercentage is kept at 0.5,

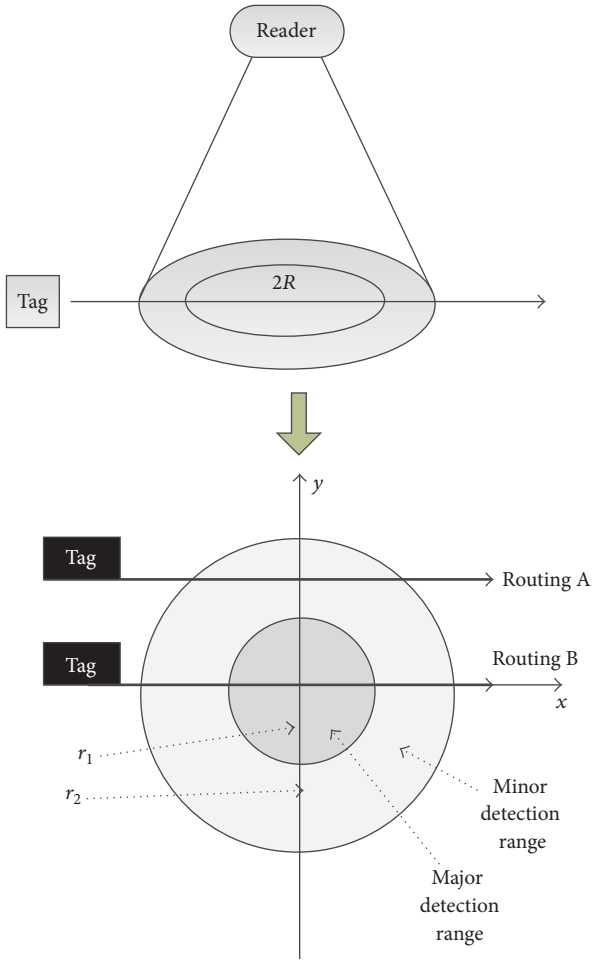


Figure 8: The experimental model of VSMURF algorithm.

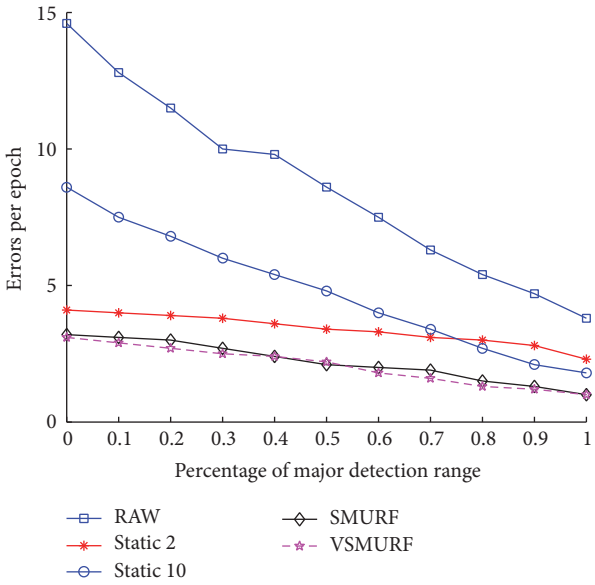


Figure 9: Results of average errors per epoch of compared algorithms (MajorPercentage is changed from 0 to 1).

Table 4: Experiment 2 parameter settings.

Parameters	Value
Detection range	20 m
Major range	10 m
MajorReadRate	0.8
Number of tags	25
Speed	Variable
Reading cycles	1000

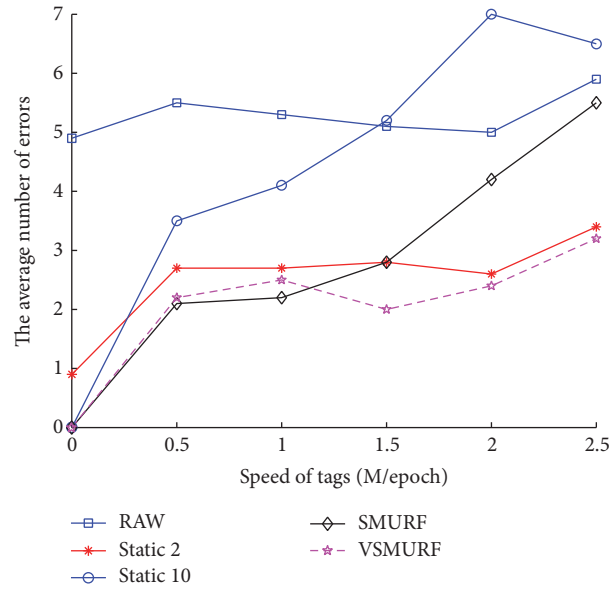


Figure 10: Results of average number of errors of compared algorithms when tags' speed changes.

MajorReadRate is 0.8, and the number of reading cycles is 1000 epochs. The fixed window cleaning algorithm, SMURF algorithm, and VSMURF algorithm are compared in this experiment. The average error rate of each algorithm is compared at different speeds. The parameter settings are shown in Table 4.

As shown in Table 5 and Figure 10, all algorithms work well when the tags' speed is less than 0.5, because the tag is in a stable environment at this speed. *Static 10*, which selects the large window to eliminate more false negative readings, obtains the best performance when the tag is static, and the number of errors per epoch is less than 1. With the tag's speed increase, the average number of errors for *static 10* starts to increase and the efficiency begins to decline. This is because the large window causes more false positive readings. In addition, the small window *static 2* cannot completely make up for false negative reading errors. However, it will not cause a lot of false positive readings, and the overall performance is better. SMURF performs well when the tag's speed is small. As the speed increases, that is, the tag moves in or out the detection range frequently, the efficiency of the algorithm is seriously degraded. VSMURF algorithm works better in most conditions and, regardless of the fact that a tag is at

Table 5: Comparison of average errors in different speeds.

Speed	Raw errors	Static 2	Static 10	SMURF	VSMURF
0	4.98	0.98	0	0	0
0.5	5.57	2.77	3.5	2.1	2.2
1.0	5.32	2.7	4.1	2.2	2.5
1.5	5.16	2.85	5.2	2.8	2
2.0	5.02	2.6	6.05	4.2	2.4
2.5	5.95	3.4	7	5.5	3.2

low or high speed, the average number of errors is the lowest in all the compared algorithms, as confirmed by the results reported in Table 5 and Figure 10.

5. Conclusion

This paper has presented a new sliding window cleaning algorithm VSMURF, which is based on the tag's dynamic property and confidence. The paper is based on the observation that the SMURF algorithm performs well only when the tag's speed is slow. However, if the tag's speed is increasing and the tag moves into or out the detection range frequently, the efficiency of the SMURF algorithm declines dramatically. To address this limitation, we have introduced VSMURF algorithm and we have shown that it performs better in most conditions and whether the tag's speed is low or high. In particular, if the velocity parameter is set to 2 m/epoch, which represents a typical situation, our proposed VSMURF performs better than SMURF. The same results apply when the tag is moving fast. The experimental results show that the proposed method is better suited than other similar algorithms in reducing the problem of false negatives. As a future work, we intend to implement VSMURF also in RFID middleware system.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Acknowledgments

This work is financially supported by Jiangsu Government Scholarship for Overseas Studies, the National Natural Science Foundation of China (no. 61373017, no. 61572260, no. 61572261, no. 61672296, and no. 61602261), the Natural Science Foundation of Jiangsu Province (no. BK20140886, no. BK20140888), Jiangsu Natural Science Foundation for Excellent Young Scholar (BK20160089), Scientific & Technological Support Project of Jiangsu Province (nos. BE2015702, BE2016185, and BE2016777), China Postdoctoral Science Foundation (no. 2014M551636, no. 2014M561696), Jiangsu Planned Projects for Postdoctoral Research Funds (no. 1302090B, no. 1401005B), Postgraduate Research and Practice Innovation Program of Jiangsu Province (SJLX15_0381, SJLX16_0326, and KYCX17_0798), Project of Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks (WSNLBZY201509), and NUPTSF (Grants no. NY214060, no. NY214061).

References

[1] T. T. Mulani and V. P. Pingle, "Internet of Things," *International Research Journal of Multidisciplinary Studies*, vol. 2, no. 1, pp. 1–4, 2016.

[2] Y. Kong, M. Zhang, and D. Ye, "A belief propagation-based method for task allocation in open and dynamic cloud environments," *Knowledge-Based Systems*, vol. 115, pp. 123–132, 2017.

[3] J. Shen, J. Shen, X. Chen, X. Huang, and W. Susilo, "An efficient public auditing protocol with novel dynamic structure for cloud data," *IEEE Transactions on Information Forensics and Security*.

[4] M. Díaz, C. Martín, and B. Rubio, "State-of-the-art, challenges, and open issues in the integration of Internet of things and cloud computing," *Journal of Network and Computer Applications*, vol. 67, pp. 99–117, 2016.

[5] J. Shen, D. Liu, J. Shen, Q. Liu, and X. Sun, "A secure cloud-assisted urban data sharing framework for ubiquitous-cities," *Pervasive and Mobile Computing*.

[6] S. F. Wamba, A. Anand, and L. Carter, "RFID applications, issues, methods and theory: a review of the AIS basket of TOP journals," *Procedia Technology*, vol. 9, pp. 421–430, 2013.

[7] J. Zhang, G. Tian, A. Marindra, A. Sunny, and A. Zhao, "A review of passive RFID tag antenna-based sensors and systems for structural health monitoring applications," *Sensors*, vol. 17, no. 2, article 265, 2017.

[8] Q. Liu, W. Cai, J. Shen, Z. Fu, X. Liu, and N. Linge, "A speculative approach to spatial-temporal efficiency with multi-objective optimization in a heterogeneous cloud environment," *Security and Communication Networks*, vol. 9, no. 17, pp. 4002–4012, 2016.

[9] C. Zhao, C. Wu, J. Chai et al., "Decomposition-based multi-objective firefly algorithm for RFID network planning with uncertainty," *Applied Soft Computing*, vol. 55, pp. 549–564, 2017.

[10] R. Derakhshan, M. F. Orlowska, and X. Li, "RFID data management: challenges and opportunities," in *Proceedings of the IEEE International Conference on RFID*, pp. 175–182, Grapevine, Tex, USA, March 2007.

[11] J. Zhu, J. Huang, D. He, Y. Leng, and S. Xiao, "The design of RFID middleware data filtering based on coal mine safety," in *Proceedings of the Second International Conference on Communications, Signal Processing, and Systems (CSPS)*, vol. 246 of *Lecture Notes in Electrical Engineering*, pp. 859–867, Springer, Tianjin, China, 2013.

[12] B. Gao, *The Research of Algorithm for Uncertain RFID Data Cleansing [Master's, thesis]*, Nanjing University of Information Science and Technology, Nanjing, China, 2015.

[13] F. Abdul Rahman, M. Desa I, A. Wibowo, and N. Abu Haris, "Antoni wibowo and norhaidah abu haris, data cleaning in knowledge discovery database (KDD)-data mining," in *Proceedings of the International Conference on Advance in Computer Science and Electronics Engineering (CSEE '14)*, pp. 4–10, New Delhi, India, February 2014.

[14] M. A. Ramadan, A. I. El Seddawy, and A. S. El Rawas, "The impact of total quality management using data mining on the integrated innovation management in smart education," *International Journal of Scientific and Research Publications*, vol. 5, no. 5, pp. 1–6, 2015.

[15] C. C. Aggarwal and J. Han, "A survey of RFID data processing," in *Managing and Mining Sensor Data*, pp. 349–382, Springer-Verlag, New York, NY, USA, 2013.

[16] Y. Liu, H. Zhang, L. Zhang, and X. Xu, "Measuring the uncertainty of RFID data streams," *Journal of Computational Information Systems*, vol. 11, no. 6, pp. 2181–2187, 2015.

[17] Z. Zhao and W. Ng, "A model-based approach for RFID data stream cleansing," in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM '12)*, pp. 862–871, New York, NY, USA, November 2012.

[18] Y. Gu, G. Yu, X.-L. Hu, and Y. Wang, "Efficient RFID data cleaning model based on dynamic clusters of monitored objects," *Journal of Software*, vol. 21, no. 4, pp. 632–643, 2010.

[19] H. Mahdin and J. Abawajy, "An approach to filtering RFID data streams," in *Proceedings of the 10th International Symposium on Pervasive Systems, Algorithms, and Networks (I-SPAN '09)*, pp. 742–746, Kaohsiung, Taiwan, December 2009.

[20] Z. Hongsheng, T. Jie, and Z. Zhiyuan, "Limitation of RFID data cleaning method—SMURF," in *Proceedings of the IEEE International Conference on RFID-Technologies and Applications (RFID-TA '13)*, 4, 1 pages, Johor Bahru, Malaysia, September 2013.

[21] Y. Bai, F. Wang, and P. Liu, "Efficiently filtering RFID data streams," in *Proceedings of the 1st International VLDB Workshop on Clean Databases (CleanDB '11)*, pp. 1–8, Seoul, Korea, September 2006.

[22] S. R. Jeffery, G. Alonso, M. J. Franklin, W. Hong, and J. Widom, "A pipelined framework for online cleaning of sensor data streams," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, pp. 1–12, Altanta Georgis, Ga, USA, April 2006.

[23] H. Gonzalez, J. Han, and X. Shen, "Cost-conscious cleaning of massive RFID data sets," in *Proceedings of the 23rd International Conference on Data Engineering (ICDE '07)*, pp. 1268–1272, Istanbul, Turkey, April 2007.

[24] Y. Wang, B.-Y. Song, H. Fu, and X.-G. Li, "Cleaning method of RFID data stream based on kalman filter," *Journal of Chinese Computer Systems*, vol. 32, no. 9, pp. 1794–1799, 2011.

[25] S. R. Jeffery, M. Garofalakis, and M. J. Franklin, "Adaptive cleaning for RFID data streams," in *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB '06)*, pp. 163–174, Seoul, Korea, September 2006.

[26] Y. Chen, C. Hao, W. Wu, and E. Wu, "Robust dense reconstruction by range merging based on confidence estimation," *Science China Information Sciences*, vol. 59, no. 9, Article ID 092103, 2016.

[27] A. F. C. Errington, B. L. F. Daku, and A. F. Prugger, "Initial position estimation using RFID tags: a least-squares approach," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 11, pp. 2863–2869, 2010.

[28] J. Dou, F. Zou, and F. Chen, "Application of the central limit theorem," *Aisa-Pacific Education*, vol. 3, no. 1, pp. 138-139, 2016.

[29] N. Dhama and M. Sharma, "An improved algorithm for reducing false and duplicate readings in RFID data stream based on an adaptive data cleaning scheme," *International Journal of Computer Trends and Technology*, vol. 4, no. 4, pp. 944–950, 2013.

Hindawi
Journal of Sensors
Volume 2017, Article ID 8782131, 15 pages
https://doi.org/10.1155/2017/8782131

Research Article

Anomaly Detection in Smart Metering Infrastructure with the Use of Time Series Analysis

Tomasz Andrysiak, Łukasz Saganowski, and Piotr Kiedrowski

Institute of Telecommunications, Faculty of Telecommunications and Electrical Engineering, University of Technology and Life Sciences (UTP), Ul. Kaliskiego 7, 85-789 Bydgoszcz, Poland

Correspondence should be addressed to Tomasz Andrysiak; andrys@utp.edu.pl

Received 10 March 2017; Revised 31 May 2017; Accepted 13 June 2017; Published 18 July 2017

Academic Editor: José R. Villar

The article presents solutions to anomaly detection in network traffic for critical smart metering infrastructure, realized with the use of radio sensory network. The structure of the examined smart meter network and the key security aspects which have influence on the correct performance of an advanced metering infrastructure (possibility of passive and active cyberattacks) are described. An effective and quick anomaly detection method is proposed. At its initial stage, Cook's distance was used for detection and elimination of outlier observations. So prepared data was used to estimate standard statistical models based on exponential smoothing, that is, Brown's, Holt's, and Winters' models. To estimate possible fluctuations in forecasts of the implemented models, properly parameterized Bollinger Bands was used. Next, statistical relations between the estimated traffic model and its real variability were examined to detect abnormal behavior, which could indicate a cyberattack attempt. An update procedure of standard models in case there were significant real network traffic fluctuations was also proposed. The choice of optimal parameter values of statistical models was realized as forecast error minimization. The results confirmed efficiency of the presented method and accuracy of choice of the proper statistical model for the analyzed time series.

1. Introduction

Smart Metering Communications Networks (SMCN) are one of the most important parts of the Smart Grid system [1]. With smart metering, not only the remote, automatic electricity meters' reading but also the customer's switching on/off is possible. The reading process can be done very often, for example, every 15 minutes per every meter. Frequent reading allows for more accurate energy consumption forecasting because of having large statistic material based on individual electricity consumption profiles (the more accurate we forecast, the more money we save).

Smart Metering Communications Network consists of last-mile networks, access networks, and a backbone network. Both backbone and access networks are realized using typical methods, that is, using IP network as a backbone and mostly GPRS technology to access it. It should be noted that these typical solutions are not the only ones. There can be other very original solutions, for example, the one described in [2]. Last-mile smart metering networks use PLC (Power Line Communications), RF (radio frequency), or a hybrid of these technologies. In this article, like in [3], the RF technology is considered. Using RF technology based on short-range devices makes the last-mile smart metering network similar to WSN (wireless sensor network). Moreover, they also use the multihop technique to expand communication range. The value of bit rates used in these networks, which is between a few to a few hundred of kbit/s, is probably the last similarity of these networks. There are two main differences between WSNs and last-mile smart metering communication networks, namely, energy issues and memory deficit. In last-mile smart metering communication network, dedicated for automatic electricity meter reading, energy issues do not exist, which is opposite to WSNs [4]. The result of the first diversity is the difference in the applied routing protocols. In WSNs, routing protocols are oriented on the balanced involvement of intermediary nodes in the process of data transferring, while in smart metering, they are oriented on reliability of data distribution and acquisition. Memory deficit in communication nodes of the smart meters is caused

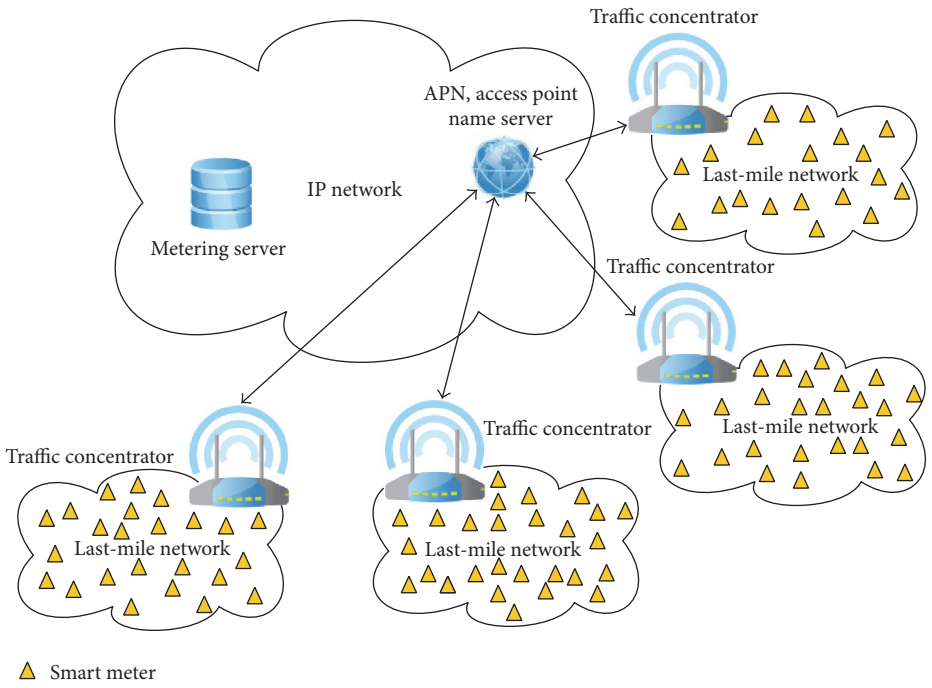


Figure 1: Smart metering critical infrastructure management system: an overview.

by using it mostly to implement sophisticated encryption algorithms, because smart metering systems, being part of the Smart Grid (which is classified as the critical infrastructure), must meet high security requirements. The result of this memory deficit forces other approaches to solve typical problems of the network maintenance. One of such problems is anomaly detection in last-mile network. It is impossible to implement even simplest anomaly detection algorithms, even though there is a spare memory, because this spare memory is reserved for the future for new more sophisticated encryption algorithms. Independent from the memory insufficiency, the second reason of difficulties with anomaly detection in smart meters is that the throughput of the last-mile network is too small to report detected anomalies in the right time. Moreover, most of the anomalies would also be detected by the neighbouring nodes, which multiplexes data traffic in the network. In smart metering, the last-mile networks operate at nearly maximum traffic load. The typical number of smart meters in a single last-mile network is around 250. Every smart meter must be read out every 15 min and it takes a few seconds (from 1 sec. to 4 sec. typically). There is only a small margin of bandwidth to support the maintenance and management or to enable the reading process during degraded propagation conditions. The above reasons induced us to carry out detection of anomalies in the data traffic concentrator. The data traffic concentrator (TC) is a thick node similar to the sink in WSNs. The construction of it is mostly based on the single-board computers which have enough RAM and ROM memory and also have a fast processor. The data traffic concentrator is connected to both last-mile and access networks. It is easy to update when there is access to network database of anomalies or the detection methods.

Bearing in mind the above, we have chosen to detect network anomalies by means of exponential smoothing of statistical models and outliers detection. The purpose of the proposed operations is to examine differences between real network traffic parameters and the same traffic's estimated statistical models. A two-stage anomaly detection method was used for the process mentioned above. Its first part consisted in seeking and elimination of any outliers in traffic parameters of the advanced metering infrastructure (AMI). This step was based on Cook's distance, which is a simple and efficient method. Consequently, in the second part of the process, the data which remained served as a base for creation of statistical models by means of exponential smoothing. In result, the operation showed differences in the tested AMI parameters.

In our solution, three types of anomalies were tested: (i) energy theft by bypassing electricity meter and energy meters shielding, (ii) electromagnetic distortion caused by Radio Frequency Interferences (RFI) and conducted interferences through power mains, and (iii) interference of communication caused by coordinated attacks.

General overview of Smart Grid advanced metering infrastructure (AMI) is presented in Figure 1. A last-mile network consists of AMI network realized by means of wireless sensor network (WSN). Power meters have built-in wireless sensors, working in industrial, scientific, and medical (ISM) bands. Traffic from power meters is received by a traffic concentrator, which plays a role of communication gateway between a WSN network and other communication links realized by, for example, IP network, General Packet Radio Service (GPRS), or Long-Term Evolution (LTE). Every traffic concentrator communicates through access point name

(APN) server (see Figure 1) which represents a link realized by packet communication network. In higher energy operator, the level application installed on the metering server is responsible for maintenance and billings.

The article is organized as follows: after Introduction, Section 2 describes communication scheme used in the last-mile test-bed network. Next, Section 3 presents related work on existing anomaly detection systems for Smart Metering Communications Network. Section 4 discusses the categories and nature of AMI security questions. Section 5 presents the structure and functioning of the research system. In Section 6, the real-life experimental setup as well as experimental results is presented. Finally, Section 7 concludes our work.

2. Communication Scheme Used in the Last-Mile Test-Bed Network

Communication scheme used in the examined last-mile network was designed by one of the coauthors in 2010 and published in 2011 in [5] as EGQF (Energy Greedy Quasi-Flooding) protocol. This paper presents only necessary information about the scheme for better understanding of the methods of anomaly detection. The EGQF protocol is independent from communication media types and may be used in networks using RF, PLC, or even RF/PLC [6] hybrid technologies. It uses the multihop technique for an extending transmission range and also the multipath technique to improve reliability of data transfer. The architecture of the presented network is very simple because it can operate having only two types of nodes: a traffic concentrator and electricity meters. The traffic is forced and coordinated by the traffic concentrator. At the same time, only one electricity meter is queried. All the other nodes, which are not queried at the moment, can act as transfer nodes relaying packets to or from the destination node. Due to the lack of memory, terminals do not know the network topology and even do not know the addresses of neighbouring nodes.

The EGQF protocol uses a reduced set of packet types, that is, command packets, response packets, and ACK/Cancel packets. Command packets, in most cases, are used by the traffic concentrator for querying the electricity meter. The response from the electricity meter is transported over the response packet. The ACK/Cancel packet is a packet which acts as the ACK for the destination node and as the reading process canceller for the other nodes. The ACK/Cancel packet can be sent only by the traffic concentrator to confirm the reception of the response and to put out the flooding of remaining response copies. The relaying process in nodes, which are neither destination nor source nodes, depends on transmitting the copy of the packet after random time in the condition of a not detected carrier. The difference between the typical flooding protocol and the EGQF protocol is that using a typical flooding protocol nodes sends a copy of packet always once during the transferring process, while when using the EGQF protocol, copies are sent as often as needed, for example, once, twice, or not at all. The decision whether a copy of the packet should be sent is made when the transfer discriminator (TD) value of a packet is greater than the previous stored one. Initial (or set at the end of the process) transfer discriminator value is zero. The transfer discriminator consists of two fields organized in the following order: the packet type code and the time to live (TTL) counter. The TTL occupied the least three significant bits of the control field of the packet, while the packet type code occupied two more significant bits in the same field, so that the transfer process of command packet is always canceled after receiving a response packet. It is the same with response packet transfer after receiving ACK/Cancel.

These two cases show us a situation when the copy is not sent, which is different with regard to the typical flooding protocol. This solution reduces the risk of collision. Using the same solution, it is possible to send the copy of the same packet type more than once. Such situation occurs when after sending the copy of the packet the same packet is received but with smaller value of TTL. This situation does not occur very often (i.e., when a packet with a greater number of hops came earlier than a packet with a smaller number of hops), and it increases reliability [6, 7].

Only the response and command packets can have payload field. Payload field is encrypted by the application layer, whereas the rest, like overhead, is transmitted in open unencrypted mode. So it is impossible to change readouts (attack the application layer), but it is possible to generate extra traffic by the extra node which has the same address as the existing, in last-mile, smart meter. Such an attack on confidentiality causes deterioration in network performance and can even make the real smart meter unreachable, for example, by sending copies of the response packets with small value of TTL.

3. Related Work

In most cases, anomaly detection in LV network depends on energy theft detection. The oldest method depends on finding irregularities from the customer billing centre [8]. This centralized method does not allow reacting quickly because of having historical long-term consumption records. Therefore, in [8], the new decentralized method based on short periods customers' consumption profiles is proposed. In [9], the authors used a variety of sophisticated techniques also for theft detection. There are a lot of works which focus on communication security by means of encryption or key distribution, for example, [10, 11].

This work focuses on anomaly detection in last-mile RF Smart Grid communication network, which is not only the result of the energy theft but also the result of deliberate, malicious customers' behavior or simply unconscious disturbing actions coming from other systems. There is a similar work [12], in which anomaly detection is realized neither in the central point nor in electricity meters but in a simple way. The proposed methods of anomaly detection presented in [12] are mostly dedicated for thefts detecting, while we focused on any anomaly detection in communication.

Anomalies in communication may be caused by various factors, for example, a human or independent of human activity and unintentional or intentional actions, such as theft, for instance. There are quite a lot of works dedicated

to anomaly detection in communication networks, also in Smart Grid communications systems [13–16], including the last-mile area of their communication networks. However, in these works, the authors focus on anomaly detection in an IP network, where also for smart metering last-mile network the data is carried over IP if PLC PRIME or G3 interface was implemented [17]. We used RF technology for last-mile network, where IP technology implementation was not possible, because it would lengthen the radio frames and make the radio transmission unreliable.

In literature, most anomaly detection systems are focused on anomalies in power distribution systems such as transmission line outages, unusual power consumption, and momentary and sustained outages [18]. In our work, we proposed anomaly/attack detection system in last-mile RF Smart Grid network (not in IP network). We proposed the two-step method of anomaly detection dedicated for last-mile RF communication network consisting of nodes, which are based on short-distance devices with the memory deficit and reduced protocol stack, that is, one protocol both for the data link layer and for the network layer.

4. Security in Smart Metering Communications Network

Ensuring security and protection of data collected by the smart metering systems is an exceptionally essential element of the SMCN solutions. It is obvious that data gathered by smart meters say much about private aspects of the recipients' lives. Moreover, having additional information such as sequences of readings, types of devices, or the number of inmates, it is easy to create a precise profile of daily living activities of the observed recipients, which in consequence may lead to serious abuses [19, 20].

The threats coming from the recipients themselves who have the smart metering infrastructure are not a less important security problem. The recipients can perform destructive activities on AMI, which consist in disturbing data saved in the meter, reconfiguration of settings and parameters of the counter, disruption of data transmission, or replacement of the internal counter's software so that it conveys understated values of consumed energy [9, 12, 21].

However, what appears to be a more serious problem is protection against cyberattacks [22]. A large-scale application of smart metering creates new entering possibilities for an unauthorized use by information systems. Joining of smart meters with information networks of energy companies, energy sellers, and companies managing distributed generation is essential for proper functioning of smart power networks. Thus, every meter becomes a potential entering point for a cyberattack [23]. Protection of smart networks against such attacks seems to be a more complex task and much more difficult to solve in comparison with ensuring security to data collected by smart meters or prevention from the users' abuses.

Cyberattacks onto the SMCN security may be divided into two elementary groups: passive and active attacks. The passive ones are all the attempts of an unauthorized access to data or the SMCN infrastructure, in which the attacker does not use emission of signals which may disturb or even disenable correct work of the system. Active attacks, on the other hand, are all the attempts of an unauthorized access by the attacker to data or the SMCN system's infrastructure with the use of emission of any signals or activities that can be detected [24–26].

While performing a passive attack onto the SMCN, the attacker disguises their presence and tries to obtain access to the transmitted data by passive monitoring of the network. For protection against such incidents, different cryptographic mechanisms are often used. Another passive form of attack onto the SMCN is activities aiming at obtaining an analysis of the traffic within the network. In this case, the attacker's intention is not acknowledging the content of transmitted data packets but is gaining knowledge about topology of the wireless sensor network. Due to the above, collecting information on the basis of traffic analysis in the SMCN gives the intruder knowledge about the network's critical nodes which ensure its proper work [25].

Contrary to the above presented passive methods of attack onto the SMCN, by using active attack forms, the intruder directly or indirectly influences the content of the sent data and/or the network's operational capabilities [26]. Attacks of this kind are easier to detect in comparison to the passive ones because they have direct impact onto the SMCN performance quality. An effect of an active attack may be, for example, degradation of services, or, in extreme cases, lack of access to particular services, or even a complete loss of control over the SMCN network.

Active attacks can be divided into three groups [25, 26]: (i) physical attacks, destruction of a node, a node manipulation, and electromagnetic pulse (EMP); (ii) attacks onto integrity, confidentiality, or privacy of data (including unauthorized access to data); (iii) attacks on services (Denial of Service (DoS) or Distributed Denial of Service (DDoS)) and attacks directed at each SMCN network layer.

The physical attacks are direct destructive operations that aim to physically destroy or damage the AMI infrastructure. A similar role can be performed by attacks using short-term high-energy electromagnetic pulse (EPM) or high pulse distortion in the supply network [27, 28].

The attacks directed at integrity or confidentiality of data are exceptionally dangerous because they enable the attacker to gain an unauthorized access to the AMI and to data transmitted by it. One of possible forms of such activity is the Sybil attack. It consists in compromising the network's legal node and the takeover of its identifier together with access to the AMI infrastructure [29].

Another type of attacks is a Wormhole attack [30]. In this case, the attacker creates additional links and transmits packets to an unauthorized node in WSN network. This type of attack may have serious impact on routing process and can be an introduction to other more serious attacks such as "man into the middle" attack. Overall network performance can also be downgraded because of inefficient resource utilization.

The DoS/DDoS attacks in the SMCN lead to an overload of the attacked nodes and thereby they disenable acquiring data from the attacked nodes or they preclude using the

services offered by the attacked network. Attacks of this kind are usually realized by introducing network traffic which is bigger than it is possible to service. They can have different characters; for example, they may appear in the physical layer and take the form of jamming, and in the data link layer they may flood the network with packets, simultaneously leading to data colliding and necessity to retransmit it. Appearance of the DoS attack in a network layer, on the other hand, may consist in sending packets in the wrong direction [24, 31].

To protect against the above-mentioned threats, in particular different kinds of active and passive attacks, it is necessary to ensure a high level of security to the SMCN infrastructure by application of the following rules concerning sending information and the used functionalities [32, 33].

Confidentiality. Data sent by means of the chosen communication standard, and in particular sensitive data, should be inaccessible to outsiders. It means that no person from outside can obtain access permissions of the consumer or service supplier and that the information recipients themselves do not have access to the sensitive data allowing performing unauthorized profiling, for example, do not have access to information about performance of particular devices but only to aggregated power consumption.

Integrity. This requirement must ensure that the received message has not been changed during transmission. In case of last-mile networks, integrity has impact on proper and not delayed data transmission. Change in the information content, as a result of interference or a hacker's attack, could cause rupture in communication and activation of the wrong device.

Authorization. This operation is used for identification of devices and nodes and verification of the source or origin of the data in the network. Authorization is essential at the level of administrative task realization in the network. What is exceptionally important is proper authorization of numerators of the AMI and particular network's devices, because it conditions correct performance of the system as a whole.

Accessibility. This concerns access to the network, even in cases of attacks and possible damage to the devices. The infrastructure should be designed in such a way that its resources, for example, computational capabilities and memory, would enable full functionality with maximum process involvement of its elements.

Time Sensitivity. Every sent piece of information, offset by a particular fixed time window, may become useless. The network must retain the ability to communicate with certain time delays. In case of home metering infrastructure, time sensitivity is connected to response time, that is, time counted from the service claim to proper receiver's response. Assurance of appropriate response time conditions proper realization of the claimed service.

The problem of advanced metering infrastructure's digital security is a complex and difficult task to realize in practice. It requires designing and introducing high efficiency mechanisms of safety and security in order to provide confidentiality and integrity of data, preventing abuse caused by recipients, as well as detection and neutralization of attacks. One of the possible solutions to so-stated issue is implementation of abnormal behavior detection system for particular SMCN parameters, which points at a possibility of a given abuse appearance.

The above-mentioned solution is the main focus of the present paper.

5. Methodology of Anomaly Detection System: The Proposed Solution

In order to ensure appropriate level of security to critical infrastructures such as Smart Metering Communications Networks, in particular AMI last-mile network, it is necessary to monitor and control those infrastructures simultaneously. Only this type of activities enables detecting and minimizing the results of different kinds of abuses, coming from the inside (unauthorized and/or destructive actions of the recipient) as well as the outside (attacks realized by cybercriminals) of the protected infrastructure [19].

The most often implemented solutions, realizing so-stated aim, are the IDS/IPS systems *(Intrusion Detection System/Intrusion Prevention System)*, that is, mechanisms of detection (IDS) and preventing intrusions (IPS), operating in real time [34]. In the hierarchy of critical infrastructure, they should be placed just after security elements, such as firewalls. IDS systems are used for monitoring threats and incidents of safety violation and for informing about their occurrence. The IPS systems, on the other hand, additionally take actions to prevent an attack, minimize its effects, or actively respond to security violation. Thus, the mentioned solutions allow for an increase in the level of protection of the AMI infrastructure by means of strengthening communication control between its different elements.

The IDS systems may be classified as belonging to one of two groups using different techniques of threat identification. The first one is based on detection of known attacks by means of defined, specific (for them) features, called signatures. The second, on the other hand, is based on an idea of monitoring the system's normal operation in order to detect anomalies, which may proclaim an intrusion [34, 35].

The basic advantage of methods based on anomaly detection is the ability to recognize unknown attacks (abuses). These methods use knowledge of not how a particular attack looks like but of what does not correspond to defined norms of the network traffic. Therefore, the IDS/IPS systems founded on the use of anomalies are more efficient and effective than systems using signatures in the process of detecting unknown, new types of attacks (abuses) [36].

Bearing in mind the above, for the purpose of this research paper, we decided to detect anomalies by means of performing an analysis of deviations from the real AMI last-mile traffic parameters with regard to the estimated statistical models (Figure 2). In our method, detecting anomalies is performed in two steps. In the first stage, three exponential smoothing models are formed as a basis for the AMI network traffic parameters. For this reason, prior to creating

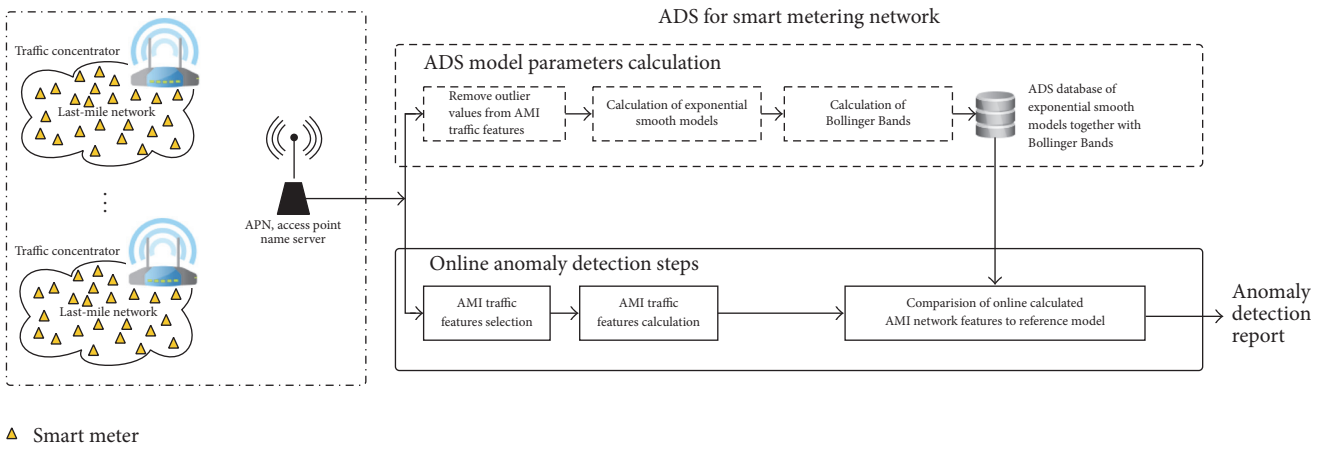


Figure 2: General overview of the proposed anomaly/attack detection method for AMI smart metering network.

the models, features of the network traffic are chosen and calculated by means of outliers detection and their exclusion. Next, the exponential smoothing models of parameters are estimated (on the basis of features of the analyzed AMI network traffic). In consequence, we obtain statistical models which serve as a basis for anomaly detection method. In the second stage, anomaly detection systems choose and estimate appropriate features of the network traffic, after which they compare the differences between the real network traffic and the calculated statistical models to perform AMI network parameters assessment.

In Figure 2, we can see a block scheme of the proposed anomaly detection method. Traffic from AMI last-mile network is captured by means of APN gateway. The proposed method is divided into two main steps. First step consists of calculation of model reference parameters (the elimination procedure of outliers' observations is realized at this stage) from extracted AMI network traffic features. Models for AMI network traffic features can be updated when the model is not up to date because of different reasons, for example, network architecture changes. Model parameters are calculated based on three different exponential smoothing models and Bollinger Bands calculation (see Sections 5.2 and 5.3). Reference models are used for comparing online the extracted AMI network traffic features in the second step of the proposed method. When calculated online, values of AMI network exceed parameters stored in the ADS reference model. The database anomaly report is detected for a given traffic feature (more explanation is presented in Section 6).

5.1. Outliers Detection and Elimination: Cook's Distance. Due to the nature of the Smart Metering Communications Networks' infrastructure (which is similar in many ways to WSN), there is a real threat of significant fluctuations of the analyzed traffic parameters in a network, that is, high likelihood of occurrence of outliers. Origin of the mentioned fluctuations may vary, for example, radio wave propagation (environmental source), changes to the infrastructure (technical source), hardware damage, an aftermath of a network attack, and intended deceit of users. Construction of a statistical model on a set of such data may lead to many unfavorable consequences. It is then highly likely that inference, predication, and decision-making process based on such a model will be burdened with big errors, and the created model will not reflect the main mechanisms regulating behavior of the analyzed phenomenon. Therefore, evaluation of influence of particular observations onto the final result should be an essential element of initial data analysis. It would allow detecting outliers and eliminating them from the data set.

In our approach, identification of outliers in the analyzed SMCN traffic parameters is performed by means of a method using Cook's Distance [37]. The essence of this method is estimation of the distance which states the level of data matching for two models: (i) a complete model, which includes all observations from the learning set, and (ii) a model built on a set of data, from which one i observation was omitted.

$$D_i = \frac{\sum_{j=1}^{n} \left(\widehat{Y}_j - \widehat{Y}_{j(i)}\right)^2}{m \cdot \mathrm{MSE}}, \tag{1}$$

where $\widehat{Y}_j$ is the forecasted value of x variable for observations number j in the complete model, that is, built on the whole learning set; $\widehat{Y}_{j(i)}$ is the forecasted value of x variable for observations number j in the model built on the set $Y_{j(i)}$, where i is number of observations that were temporarily deactivated, MSE is the mean-model error, and m is the number of parameters used in the analyzed model.

For Cook's distance D_i threshold value, above which the given observation should be treated as an outlier, in compliance with criterion (1), 1 is accepted, or alternatively

$$\frac{4}{n - m - 2}, \tag{2}$$

where n is the number of observations in the learning set.

5.2. The Exponential Smoothing Models for Estimation of AMI Traffic Features Value. The exponential smoothing methods

are a wide range of statistical models with different assumptions and complexity levels, which emerge from a common idea of creating forecasts by means of weighted moving averages. The common denominator of those methods is assigning (exponentially) weight decreasing with distance in time to past observations in the process of setting new forecast of a future observation [38].

It is easy to notice that exponential smoothing models are based on a sensible assumption that the future value depends on not only the last observed value but also their whole series of the past values. At the same time, the influence of old values (previous) is smaller than the influence of the new values [39].

Great practical importance of exponential smoothing models is based on the fact that they are suitable for forecast construction not only in conditions of stabilized development of phenomena to our interest but also when this development is irregular, characterized by trend's fluctuations. In these models, solid analytic trends are not accepted. To the contrary, it is assumed that, for every period, assessment of the trend's level and possible periodical fluctuations are built as some average from these kinds of evaluations made in previous periods [38, 40]. Among many representations known in literature, in this paper, the following models will be used: Davies and Brown [41], Holt's linear [42], and Winters' [43] models. It is due to a different representation of the compositional models of the analyzed time series and willingness to determine possibly the best model for the presented method of anomaly detection.

5.2.1. Brown's Model. A simple model of exponential smoothing, otherwise called Brown's model [41], is one of the methods most often used in case of a time series with fixed or very weak trend, when the series does not show developmental trend and fluctuations of its values result from random factors. This method consists in smoothing the time series of the forecasted variable by means of weighted moving average; however, the weights are defined according to exponential rule.

This model can be described by means of the following recurrent formula:

$$F_1 = x_1, \tag{3}$$

$$F_t = \alpha F_{t-1} + (1-\alpha) F_{t-1}, \tag{4}$$

where $x_1, x_2, \ldots, x_n$ are values of the forecasted series, F_t is the value of the forecast in time t, and α is a parameter of the model, so-called smoothing constant, with the value of $\alpha \in [0, 1]$.

The conclusion from (4) is that the value of forecast in time t depends, in recurrent manner, on the value of the time series and forecasts for times $t-1, t-2, \ldots, 1$. As the value of forecast F_1, necessary for construction of the model, we most often accept the initial value of the variable forecasted in the time series, that is, x_1, or arithmetic average of few first values of the variable x from the time series.

The value of coefficient α influences the degree of a time series smoothing, so if $\alpha \approx 1$, then the constructed forecast will highly count the ex post errors of the previous forecasts. However, in the opposite case, when $\alpha \approx 0$, the built forecast will employ those errors to a small extent. Brown assumed that the parameter α should equal $2/(n+1)$, where n is the number of observations [44].

Because the size of coefficient α has impact on the quality of the predictive model and the size of forecasts' errors, it is impossible to point arbitrarily the best value of that coefficient for every data. Therefore, this problem can be defined as an optimization task; that is, we are looking for such an $\widehat{\alpha}$, for which

$$s(\widehat{\alpha}) = \min_{\alpha \in [0,1]} s(\alpha), \tag{5}$$

where $s(\alpha)$ denotes an objective function, which characterizes the standard forecast error.

The often used objective function is

$$s(\alpha) = \frac{1}{n} \sum_{t=1}^{n} \left| F_t - x_t \right|, \tag{6}$$

which describes mean absolute forecast error. Its form is essential, because minimization of the objective function (5) is minimization of the sum of absolute deviations. This problem is easy to check for computationally simpler linear programming problem.

5.2.2. Holt's Linear Model. For smoothing and forecasting a time series, in which developmental model and trend of random fluctuations may be present, Holt's model [42] is most often used. It is described by means of two parameters, α and β, and it then takes the following form:

$$\begin{aligned} F_1 &= x_1, \\ S_1 &= x_1 - x_0, \\ F_t &= \alpha x_1 + (1-\alpha)\left(F_{t-1} + S_{t-1}\right), \\ S_t &= \beta\left(F_t - F_{t-1}\right) + (1-\beta) S_{t-1}, \end{aligned} \tag{7}$$

where $x_1, x_2, \ldots, x_n$ are the values of the forecasted series, F_t is the smoothed value of the time series, S_t describes the smoothed trend's growth value in the moment of time t, variables α and β are the model's parameters, and t indexes the consecutive time moments.

The values of F_t and S_t are calculated in recurrent manner. The forecasts of the future time series' values, however, are determined in the following way:

$$x^*_{n+k-1} = F_{n-1} + k \cdot S_{n-1}, \quad k = 1, 2, 3, \ldots. \tag{8}$$

Holt's model's parameters α and β are chosen in such a way that they minimize possible errors of the expired forecasts. For this reason, specific values of these parameters are taken and determined, in compliance with dependency (8), with the assumption that $n = t$ and $k = 1$ are the expired forecasts.

$$x^*_t = F_{t-1} + S_{t-1}, \tag{9}$$

for time moments t, where $t = 2, 3, \ldots, n-1$ on the basis of series values from the previous period $\{x_1, x_2, \ldots, x_{t-1}\}$. These forecasts can be compared to factual values of the series x_t. The obtained differences are errors of the expired forecasts which are given by a model for the taken parameters α and β. As a measurement of the method's quality, the average of errors of the expired forecasts should be acknowledged. It may be a linear average,

$$J_1 = \frac{1}{n-2}\sum_{t=2}^{n-1}\left\lceil F_{t-1} + S_{t-1} - x_t \right\rceil, \tag{10}$$

or root mean square,

$$J_2 = \sqrt{\frac{1}{n-2}\sum_{t=2}^{n-1}\left(F_{t-1} + S_{t-1} - x_t\right)^2}. \tag{11}$$

Finally, it is necessary to choose from all possible α and β parameter values such data that provides the lowest error value J_1 or J_2. By doing so, optimal parameters values and a measure of the forecast error are determined for the taken model. It is commonly accepted that $\alpha \in [0, 1]$ and $\beta \in [0, 1]$.

5.2.3. Winters' Model. Winters' model is a generalized Holt's model form. It is used for forecasting and smoothing a time series, in which a seasonal component, development trend, and random fluctuations may occur. There are two most often used types of Winters' model: (i) multiplicative model, when the level of seasonal fluctuations around the trend increases or decreases (more precisely when the relative level of seasonal fluctuations is approximately constant), and (ii) additive model, when the level of seasonal fluctuations around a weak trend or a constant level does not change, that is, when the absolute level of seasonal fluctuations is approximately constant. For the purpose of the presented solution, only the additive model will be described and used.

Winters' [43] model is described by means of three parameters, α, β, and γ, representing, respectively, the smoothing constant for the trend's level, the change in the trend's level, and seasonal fluctuations. For so-described parameters, it then takes the following form:

$$\begin{aligned} F_{t-1} &= \alpha\left(x_{t-1} - C_{t-1-r}\right) + (1-\alpha)\left(F_{t-2} + S_{t-2}\right), \\ S_{t-1} &= \beta\left(F_{t-1} - F_{t-2}\right) + \left(1-\beta\right)S_{t-2}, \\ C_{t-1} &= \gamma\left(x_{t-1} - F_{t-1}\right) + \left(1-\gamma\right)C_{t-1-r}, \end{aligned} \tag{12}$$

where $x_1, x_2, \ldots, x_{n-1}$ are values of the forecasted series, F_{t-1} is the smoothed value of the forecast variable in moment $t-1$ after elimination of the seasonal values, S_{t-1} describes with evaluation the increment trend in the moment of time $t-1$, C_{t-1} is evaluation of the seasonal index in the moment $t-1$, r is the length of the seasonal cycle (the number of phases in the cycle, where $1 \le r \le n$), variables α, β, and γ are the model's parameters with values from the range [0, 1], and t is an index of the following moments of time.

The forecast x_t^* in the moment of time t is given by the following dependency:

$$x_t^* = F_n + S_n\left(t-n\right) + C_{t-r}, \quad t > n. \tag{13}$$

Parameters α, β, and γ are chosen similarly as in Holt's model, minimizing the mean square error of the expired forecasts; or values close to 1 are chosen when the components of the time series change quickly; or values close to 0 are chosen when the series' components do not show quick changes.

As values of F_1, S_1, and C_1, we take, respectively, the value from the time series corresponding to the first phase of the second cycle (or the average value from the first cycle), the difference of the average values from the second and first cycles, and the quotient value of the variable in the first cycle in relation to the average value in the first cycle.

5.3. Estimation of the Forecast Variability: Bollinger Bands. Bollinger Bands is a tool of technical analysis invented by Bollinger at the beginning of the 80s of the 20th century [45]. It was created on the basis of observation of financial instruments volatility. It is composed of three elements: (i) the middle band (core), which is n periodic moving average; (ii) the upper band, being k times of n periodic standard deviation above the middle band; and (iii) the bottom band, being k times of n periodic standard deviation below the middle band. The main idea of this tool is the rule that when data variability is low (their standard deviation decreases), then the bands shrink. However, in case the data variability increases, the bands expand. Thus, this tool shows dynamics of data variability. It usually defaults to the values of parameters $k = 2$ and $n = 20$ [46]. Such approach is based on the assumption that, in data of normal distribution, the area of two standard deviation widths includes 95 percent of all observations.

In the presented solution, we used Bollinger Bands to estimate forecasts variability of the exploited statistical models. As the middle band (the core), we adopted the values of statistical models' forecasts, k was the double standard deviation, and $n = 15$ (due to the 15-minute analysis windows). Figure 3 presents an exemplary PPM signal and Bollinger Bands created on its base (for Holt's model).

5.4. The Condition of Model's Parameters Update. It is possible that data in the analyzed time series will fluctuate due to the nature of the AMI network traffic parameters. The reasons for such a phenomenon are to be found in possible changes of the AMI network infrastructure (ageing of devices and replacement with new/other models) or emergence of permanent obstacles, which have significant impact on the transmitted radio signal. These factors should cause adapting of the proposed anomaly detection method to the changing conditions (which are not an aftermath of any abuses). One of the possible solutions to so-stated problem can be an update procedure of the reference statistical models, realized on new data sets which contain the subject fluctuations.

The condition for creation of a new reference model should be detection of a significant and possibly permanent statistical variability in the analyzed data set (elements of a time series). Assuming a close-to-normal data distribution, we can deduce that in the range of width of six standard deviations there is over 99 percent of data. Thus, if we define

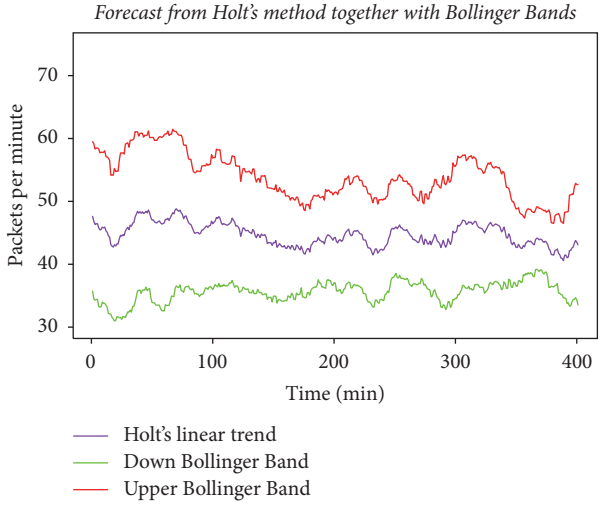


Figure 3: Exemplary Bollinger Bands for packets per minute (PPM) network feature.

the average on the basis of the forecast set of the given exponential smoothing model, and the standard deviation is estimated for the real values of the analyzed data, then a great degree of not fulfilling the above stated condition may proclaim that the statistical nature of the analyzed data has changed.

Due to the above, the following condition can be formulated. If it is not satisfied, the reference model should be updated.

$$x_i \in (\mu - 3\sigma, \mu - 3\sigma) \quad i = 1, 2, \ldots, n, \tag{14}$$

where $\{x_1, x_2, \ldots, x_n\}$ is a time series limited by n-elements analysis window, μ is the average calculated on the forecasts of the given reference model in the analysis window, and σ is the variance of the tested time series elements in relation to such an average.

In result of conducting many experiments in the presented solution, we adopted the size of analysis window $n = 15$ and an assumption that only not satisfying condition (14) in over 30% of analysis windows in a time period of a week causes an effect in the form of reference model update.

6. Experimental Installation and Results

Figure 2 presents general overview of the proposed anomaly detection method. Traffic from 70 power meters distributed across eight buildings is captured by APN gateway through an IP link. The proposed method is divided into two mains steps: calculation of ADS model parameters and online anomaly detection. In both steps, we have to extract AMI traffic features proposed in Table 1. After that, we calculate initial reference models for every traffic feature. Models are calculated for a period of one week and time is divided into 15 minutes' analysis windows. Every traffic feature is organized as one-dimensional time series. First substep in model parameters calculating removes outlier values (see Section 5.1) from every traffic feature in order to remove

Table 1: AMI network traffic features captured from sensor network gateway.

Network feature	AMI network traffic feature description
NF_1	RSSI: received signal strength indication for AMI power meter [dBm]
NF_2	LQI: link quality indicator value (values: 0–127)
NF_3	PER: packet error rate per minute [%]
NF_4	PPM: number of packets per minute
NF_5	TTL: packet time to live value

suspicious values from the model calculation. After that, we calculate exponential smooth models with the use of three exponential smooth models: Brown, Winters, and Holt (see Section 5.2).

In the next step, we compute Bollinger Bands (see Section 5.3) for achieving network traffic features variability intervals. In the end, we save models parameters together with associated Bollinger bands to database of reference models. In the second step of the proposed method, we compare values of online extracted AMI network features to reference models stored in the ADS database. ADS model gives us variability interval/variability canal for a given traffic feature. When the online calculated AMI traffic features values do not exceed interval set by the reference model, we assume that there is no anomaly/attack for a given traffic feature. When network traffic exceeds values set by the reference model, an anomaly detection report is generated for a given traffic feature.

The method proposed so far would not be resistant to AMI network changes, like increasing number of sensors or topology changes. In these cases, the reference models will not be updated and the number of FP indicators would increase in time. That is why we propose a trigger condition which is responsible for initiation of model parameters recalculation (see (14)). When the proposed condition is not satisfied in 30% of 15 minutes' analysis windows (30% of analysis windows in a period of one week), we recalculate traffic profiles for a period of one week (network traffic values are always stored for a period of one week which is why we can always recalculate traffic profiles when condition from (14) is not satisfied). New ADS network profiles are always active since the beginning of a new week.

6.1. Experimental Setup and Results. In this section, we showed experiments and results obtained in real-world test of the AMI power meter network. We proposed four different scenarios that trigger anomaly/attack in our test network. We proved that the proposed anomaly/attack detection method can be useful in detection of unwanted situations in the AMI measurement network.

The anomaly detection method presented in the article was evaluated by means of real-world installation of AMI network. The AMI network traffic was captured from installation placed in our university building [47]. The network consisted of 70 sensor nodes installed within energy power meters (see Figure 4). Sensors were installed on four floors (see Figure 6),

Table 2: DR [%] and FP [%] for anomalies/attacks performed on AMI network with Scenario 1.

Feature	Holt		Winters		Brown		Description
	DR [%]	FP [%]	DR [%]	FP [%]	DR [%]	FP [%]	
NF_1	92.40	8.80	90.20	9.80	88.10	10.40	Significant impact on NF_1 in Scenario 1
NF_2	96.00	5.60	94.10	7.50	90.20	9.80	Significant impact on NF_2 in Scenario 1
NF_3	91.00	9.40	88.40	11.30	86.30	12.70	—
NF_4	81.40	9.20	89.10	11.10	86.20	12.60	—
NF_5	72.20	10.20	70.10	12.60	68.20	12.80	—

Figure 4: Opened power meter with signed WSN communication radio module.

Figure 5: Cluster of electricity power meters in building 2.3.

located in eight separate buildings. In Figure 5, we can see a cluster of electricity meters installed in building 2.3 (see Figure 6). A traffic concentrator was placed on the second floor. Traffic from the AMI network was captured from IP connection of the traffic concentrator signed by red octagon located in building number 2.4 (see Figure 6). In the next step, we extract five traffic features NF_1–NF_5 (Table 1), where every traffic feature is represented by one-dimensional time series values.

We used these traffic features for anomaly/attack detection by means of the proposed statistical algorithm.

First two features describe the quality of the radio link: NF_1 RSSI [dBm] (higher value stands for better signal's strength) and NF_2 LQI value (values change from 0 to 127; lower values indicate higher link quality). LQI characterizes strength and quality of the received packet (in other words, how easily the received signal can be demodulated), contrary to RSSI, which gives us information about the received signal strength (it is not the measure of ability to decode signal), where radio frequency power can originate from arbitrary source such as Gaussian Frequency Shift Keying (GFSK), other ISM systems, Wi-Fi, Bluetooth, or background radiation. NF_3 and NF_4 features give us two values in a period of one minute: packet error rate (PER) per minute (number of corrupted packets received by concentrator)/(number of all packets received by the concentrator) in time interval (in our case, it was one minute) and PPM, number of packets per minute. NF_5 carries information about TTL value of packets received by a traffic concentrator. The proposed anomaly detection method was designed especially for data link and network layers. Because of security restrictions, we do not have access to the application layer payload. Application layer data is, in our case, available only for the energy supplier. We focused on detection of anomalies/attacks in layer 2 and layer 3, because there are not many anomaly detection solutions that work in last-mile AMI network. Additionally, predictable amounts of traffic made it a great candidate for anomaly detection, and we use this feature. Traffic is actually small taking into account computing power of the traffic concentrator processor but it is also large enough not to implement anomaly detection in smart meter.

We created different anomaly and attack scenarios for anomaly detection in the AMI network, and we selected four of them to evaluate the proposed method:

Scenario 1. Radio Frequency Interferences (RFI) and conducted interferences through power mains and Electromagnetic Interferences (EMI).

Scenario 2. Existence of natural and human-made obstacles,

Scenario 3. Power meter intentional damages,

Scenario 4. Coordinated attacks on power meter AMI network.

Scenarios used for anomaly/attack detection have various impacts on AMI network traffic features proposed in Table 1. In Scenario 1, we consider distortions caused by, for example, different radio ISM systems, and conducted EMI distortions carried by physical power line. A conducted EMI distortion may come from devices connected to power mains like electric engines, switching power supply, welding machines, or any industrial environment. Parts of conducted EMI distortions are presented in IEC standard 61000-4-4 [48]. We simulated some distortions that belong to both groups.

Distortions from Scenario 1 have biggest impact on network features NF_1 (RSSI) and NF_2 (LQI). Detection rate and false positive partial results for Scenario 1 are presented in Table 2.

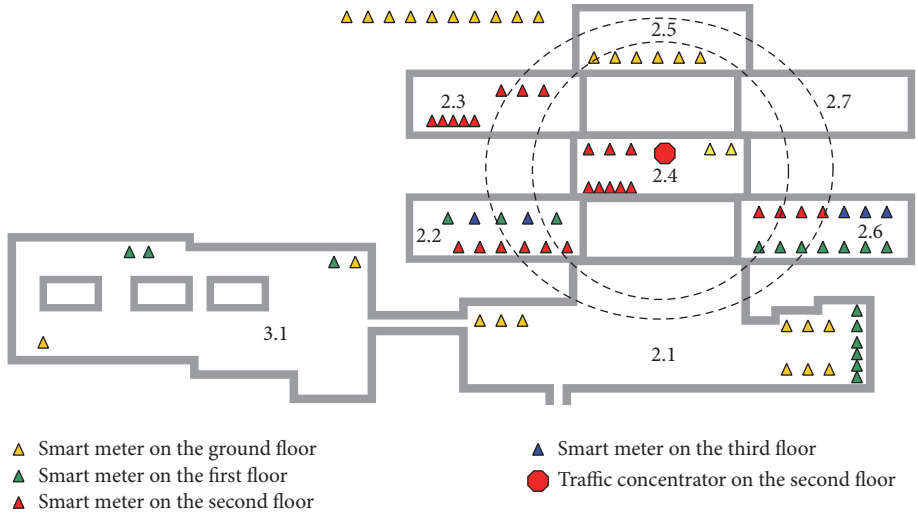


Figure 6: Physical layout of power meters of AMI network in the university building [5].

Table 3: DR [%] and FP [%] for anomalies/attacks performed on AMI network with Scenario 2.

Feature	Holt		Winters		Brown		Description
	DR [%]	FP [%]	DR [%]	FP [%]	DR [%]	FP [%]	
NF_1	88.20	8.20	86.10	10.10	83.20	11.80	Significant impact on NF_1 in Scenario 2
NF_2	92.40	5.20	90.30	7.30	87.50	9.40	Significant impact on NF_2 in Scenario 2
NF_3	82.20	9.60	80.20	11.20	76.40	12.40	—
NF_4	80.20	10.10	78.10	12.10	76.30	12.60	—
NF_5	85.60	12.20	82.40	12.60	79.40	12.80	—

An attack, according to Scenario 1, is easy to carry out, for example, by using amateur shortwave radio set to the same frequency as the working channel; modulation type does not matter. The best results in attacking give the transmitter localized close to the traffic concentrator or a cluster of electricity power meters.

Scenario 2 was simulated by locating groups of power meter sensors on different floors and distant buildings (see Figure 6). Temporarily placed obstacles, like a big truck, can also have an impact on WSN network transmission. Localization and distance between the AMI power meter sensors have impact on every captured network traffic feature. Partial results for Scenario 2 can be observed in Table 3.

The easiest way to carry out the attack according to Scenario 2 is grounding the concentrator antenna or slightly unscrewing it. In our experiments, we achieved this effect by reducing transmitting power and increasing the receiver's sensitivity simultaneously.

Intentional damage from Scenario 3 is caused by power meter users who want to avoid/delay paying electricity bills or want to bypass power meter or disturb AMI network operation. Electromagnetic metallic shielding and bypassing of power meter are exemplary methods for disturbing of the AMI sensor operation. Partial results for this scenario are presented in Table 4. Intentional damage can be seen especially for NF_3, where PER for a given power meter increases.

In our experiments, we simply turned smart meters off from mains or remotely changed the radio channel frequency just to make communication impossible. In real situation, instead of power meter intentional damaging, the easiest way to achieve the same effect is forcing the fuse protection (before input connector) to act.

Scenario 4 takes into account coordinated attacks/anomalies performed on power meters Smart Grid infrastructure. We simulate WSN flooding attack [49] and after that we add some intermediate sensor in order to perform additional links (Wormhole-type attack [30]). This type of attack/anomaly has the biggest impact on NF_4 PPM (number of packets per minute) and NF_5 TTL (packet time to live) value. Subsequent partial result can be seen in Table 5. In this scenario, traffic features (NF_1–NF_3) did not give us usable information for anomalies detection, so they can be omitted in this case.

Attacks, according to Scenario 2, were emulated by us with the use of smart meter service terminal, which is a mobile, specific kind of the traffic concentrator. We sent from service terminal to all power meters a "set date & time" command in broadcast flooding mode every 5 seconds.

Attacks described in Scenarios 1–3 require physical access, for example, in case of EMI distortions conducted through power mains or enough proximity to a selected part of physical infrastructure and in case of EMI distortions conducted through radio. Power meter shielding also requires

Table 4: DR [%] and FP [%] for anomalies/attacks performed on the AMI network with Scenario 3.

Feature	Holt		Winters		Brown		Description
	DR [%]	FP [%]	DR [%]	FP [%]	DR [%]	FP [%]	
NF_1	86.40	8.60	84.10	10.30	80.70	12.60	—
NF_2	88.40	8.40	85.20	9.80	83.10	11.70	Significant impact on NF_2 in Scenario 3
NF_3	90.50	6.40	87.20	8.80	85.60	10.90	Significant impact on NF_3 in Scenario 3
NF_4	82.30	11.50	79.50	12.40	76.40	12.80	—
NF_5	86.20	12.40	83.40	12.50	80.80	12.80	—

Table 5: DR [%] and FP [%] for anomalies/attacks performed on AMI network with Scenario 4.

Feature	Holt		Winters		Brown		Description
	DR [%]	FP [%]	DR [%]	FP [%]	DR [%]	FP [%]	
NF_1	—	—	—	—	—	—	Insignificant/negligible for Scenario 4
NF_2	—	—	—	—	—	—	Insignificant/negligible for Scenario 4
NF_3	—	—	—	—	—	—	Insignificant/negligible for Scenario 4
NF_4	92.40	6.50	90.20	8.60	87.10	10.40	—
NF_5	90.50	7.60	87.30	9.80	85.50	11.70	—

physical access to power meter. In case of Scenario 4, for example, flooding attacks on last-mile network can be performed remotely by a GPRS/IP gateway.

The anomaly detection method based on network profiles has a weakness coming from the fact that profiles are aging. This can cause an increase in the false positive (FP) values. To alleviate this effect, we propose in Section 5.4 a condition that triggers recalculation of WSN network profiles. However, there can still be situations when temporary detection rates and false positive values can be a little bit worse between the profiles' update processes. These situations may appear when we rapidly change the network structure, for example, by adding entire streets with large number of new power meters. The proposed trigger will indicate the need to recalculate new profiles, but it will happen with a programmed delay.

In order to decrease effectiveness of the proposed anomaly detection solution, the attacker needs knowledge about anomaly detection algorithms used for profiles calculation, when the system recalculates profiles, and what kinds of traffic features are extracted from the network traffic. The attacker armed with such knowledge can temporarily disturb AMI network operation between recalculations of new profiles. If the attacker has information about traffic features used by anomaly detection algorithm, he can perform an attack that would not have an impact on the proposed traffic features.

Taking into account all four scenarios, the overall performance of the proposed anomaly/attack method for five AMI network features is presented in Table 6. Most simulated attacks and anomalies were detected. In case of DR [%], values change from 68.20 to 92.26%, while FP varies between 6.40 and 12.80%. The best results for three simulated scenarios (Scenarios 1–3) were obtained for features NF_1 and NF_2. For these scenarios, features NF_1 and NF_2 were the most universal. For Scenario 4, NF_4 and NF_5 features fit better to the characteristic of simulated anomalous events. From the three evaluated models, we achieved the best results for Holt's exponential smoothing model, where not only exponential smoothing but also forecasting for time series with trend is possible.

Table 6: Overall DR [%] and FP [%] for anomalies/attacks performed on AMI Smart Grid network.

Feature	Holt		Winters		Brown	
	DR [%]	FP [%]	DR [%]	FP [%]	DR [%]	FP [%]
NF_1	89.00	8.53	86.80	10.07	84.00	11.60
NF_2	92.26	6.40	89.87	8.20	86.93	10.30
NF_3	87.90	8.47	85.27	10.43	82.77	12.00
NF_4	84.07	9.32	84.23	11.05	81.50	12.10
NF_5	83.62	10.60	80.80	11.88	78.48	12.53

Anomaly detection prediction based on Holt's exponential smoothing model gives us DR [%] values within 83.62–92.26% interval and FP [%] values changing from 6.40 to 10.60%. We were able to detect all performed anomalies/attacks described in the proposed scenarios taking into account all extracted traffic features (it was not possible to detect all anomalies/attacks by means of one traffic feature). In literature, there are many various anomaly detection methods using different algorithms [36, 50, 51] applied to WSN networks. On the basis of literature analysis, we can state that in general for WSN anomaly detection solutions FP [%] values are generally less than 10% [36, 50, 51]. Taking into account Holt's exponential smoothing model, we achieve FP values changing from 6.40 to 10.60%, so we can state that this interval is acceptable for anomaly detection class security systems.

7. Conclusions

Providing an adequate security and protection level of data sourced by intelligent measuring systems is currently an intensively examined and developed question for the world's

leading seats of learning. It is obvious that the AMI networks, due to their nature, are exposed to a significant number of threats originating from both outside and inside of their own infrastructure. Data collected recurrently by intelligent meters contain much information about private aspects of recipients' lives, which may be used for realization of serious abuse. Other, but not less important, problems of security within the AMI infrastructure are dangers coming from the recipients themselves. In some cases, they may perform actions which are destructive for the AMI. Such activities may consist in disturbing data saved in the meter or hampering their transmission. However, the key security problem is providing an adequate level of protection against external abuse, that is, safety from cyberattacks. In this case, every element of the SMCN infrastructure, AMI in particular, may become a potential attack point.

Growing level of complexity, globalization of range, and dynamically increasing number and nature of new attacks impose a change in approach towards realization of network security systems. Currently, most often implemented mechanisms are the methods of detection and classification of abnormal behaviors reflected in the analyzed network traffic parameters. An advantage of such solutions is protection against attacks unknown so far, often directed towards defined resources of critical infrastructures, or simply being the so-called zero-day exploits. Anomaly detection systems, in those cases, may play the key role. Their task is then detection (for the purposes of automatic response) of not typical behaviors in the network traffic which constitute symptoms of diverse abuse, originating both inside and outside the secured infrastructure.

The article presents an effective solution to the problem of anomaly detection in the network traffic for the critical measurement infrastructure. The structure of the AMI network, built for the purpose of the experiment, is presented and described. Crucial security problems which have a direct impact on proper operation of the advanced measurement infrastructure are discussed. A two-stage method was proposed for anomaly detection in the examined sensory network traffic, represented by proper time series. In the first stage, any possible outlying observations in the analyzed time series were detected and eliminated. The purpose of such operation was to prepare correct data for creation of standard statistical models based on exponential smoothing. Estimation of possible fluctuations of models' forecasts was realized by means of suitably parameterized Bollinger Bands. An update procedure was also proposed for the standard models in case serious fluctuations appear in the real network traffic. The second stage consisted in examining statistical relations between the standard traffic model and its real variability in order to detect abnormal behavior, which could signify an attempt of some abuse, for example, a network attack.

In the article, we proposed a method for anomaly/attack detection in data link and network layers. We did not analyze application layer, because in our case the application layer payload is only available for energy supplier. We focused on layer 2 and layer 3 because there are not many anomaly detection solutions in this area.

The proposed method of anomaly detection was evaluated with the use of real AMI network, which consists of 70 power meter nodes, located in eight distant buildings. After network traffic features extraction, we checked three different statistical models based on exponential smoothing together with Bollinger Bands. On the basis of four practical scenarios, we can conclude that the most promising results were achieved for Holt's exponential smoothing model. The proposed model fits to the characteristic of the network traffic features extracted from the AMI network. In case of Holt's model, not only is exponential smoothing possible, but also we can forecast time series with trend. We also propose a solution for aging reference models. We propose a condition (see (14)) for triggering recalculation of model parameters.

For future work, we are planning to examine usability of statistical models for anomaly detection in AMI power meter network using Power Line Communication (PLC) module instead of radio communication. In the next step, we would like to propose anomaly detection solution for hybrid AMI power meter network using at the same time radio communication and PLC communication modules.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Centre for Research and Development and by the National Fund for Environmental Protection and Water Management under the realized GEKON program (Project no. 214093), and it was also supported by the Polish Ministry of Science and High Education and Apator S.A. Company under Contract 04409/C.ZR6-6/2009.

References

[1] S. Finster and I. Baumgart, "Privacy-aware smart metering: a survey," *IEEE Communications Surveys and Tutorials*, vol. 17, no. 2, pp. 1088–1101, 2015.

[2] B. E. Bilgin, S. Baktir, and V. C. Gungor, "Collecting smart meter data via public transportation buses," *IET Intelligent Transport Systems*, vol. 10, no. 8, pp. 515–523, 2016.

[3] P. Kulkarni, S. Gormus, Z. Fan, and B. Motz, "A mesh-radio-based solution for smart metering networks," *IEEE Communications Magazine*, vol. 50, no. 7, pp. 86–95, 2012.

[4] J. N. Al-Karaki and A. E. Kamal, "Routing techniques in wireless sensor networks: a survey," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–28, 2004.

[5] P. Kiedrowski, B. Dubalski, T. Marciniak, T. Riaz, and J. Gutierrez, "Energy greedy protocol suite for smart grid communication systems based on short range devices," in *Image Processing and Communications Challenges 3*, vol. 102 of *Advances in Intelligent and Soft Computing*, pp. 493–502, Springer, Berlin, Germany, 2011.

[6] P. Kiedrowski, "Toward more efficient and more secure last mile smart metering and smart lighting communication systems

with the use of plc/rf hybrid technology," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 675926, 9 pages, 2015.

[7] P. Kiedrowski, "Errors nature of the narrowband plc transmission in smart lighting LV network," *International Journal of Distributed Sensor Networks*, vol. 2016, Article ID 9592679, 9 pages, 2016.

[8] C. Liao, C.-W. Ten, and S. Hu, "Strategic FRTU deployment considering cybersecurity in secondary distribution network," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1264–1274, 2013.

[9] S. Amin, G. A. Schwartz, A. A. Cardenas, and S. S. Sastry, "Game-theoretic models of electricity theft detection in smart utility networks: providing new capabilities with advanced metering infrastructure," *IEEE Control Systems*, vol. 35, no. 1, pp. 66–81, 2015.

[10] X. D. Wang and P. Yi, "Security framework for wireless communications in smart distribution grid," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 809–818, 2011.

[11] J.-L. Tsai and N.-W. Lo, "Secure anonymous key distribution scheme for smart grid," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 906–914, 2016.

[12] A. K. Marnerides, P. Smith, A. Schaeffer-Filho, and A. Mauthe, "Power consumption profiling using energy time-frequency distributions in smart grids," *IEEE Communications Letters*, vol. 19, no. 1, pp. 46–49, 2015.

[13] Y. Guo, C.-W. Ten, S. Hu, and W. W. Weaver, "Preventive maintenance for advanced metering infrastructure against malware propagation," *IEEE Transactions on Smart Grid*, vol. 7, no. 3, pp. 1314–1328, 2016.

[14] R. Berthier, D. I. Urbina, A. A. Cárdenas et al., "On the practicality of detecting anomalies with encrypted traffic in AMI," in *Proceedings of the 2014 IEEE International Conference on Smart Grid Communications, SmartGridComm 2014*, pp. 890–895, Venice, Italy, November 2014.

[15] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Distributed internet-based load altering attacks against smart power grids," *IEEE Transactions on Smart Grid*, vol. 2, no. 4, pp. 667–674, 2011.

[16] A. Giani, E. Bitar, M. Garcia, M. McQueen, P. Khargonekar, and K. Poolla, "Smart grid data integrity attacks," *IEEE Transactions on Smart Grid*, vol. 4, no. 3, pp. 1244–1253, 2013.

[17] ITU-T Recommendation G.9904 (10/2012): Narrowband orthogonal frequency division multiplexing power line communication transceivers for PRIME networks, 2013.

[18] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–11, 2017.

[19] V. Aravinthan, V. Namboodiri, S. Sunku, and W. Jewell, "Wireless AMI application and security for controlled home area networks," in *Proceedings of the 2011 IEEE PES General Meeting: The Electrification of Transportation and the Grid of the Future*, Detroit, Mi, USA, July 2011.

[20] Trilliant, White Papers, The Home Area Network: Architectural Considerations for Rapid Innovation, pp. 1–7, 2010.

[21] M. Balakrishnan, *Security in Smart Meters, Document number: SEC s. MTMTRWP REV0*, Free scale Semiconductor, Arizona, Ariz, USA, 2012.

[22] B. J. Murrill, E. C. Liu, and R. M. Thompson, "Smart meter data: Privacy and cybersecurity," *Smart Meters and the Smart Grid: Privacy and Cybersecurity Considerations*, pp. 1–45, 2012.

[23] V. Navda, A. Bohra, S. Ganguly, and D. Rubenstein, "Using channel hopping to increase 802.11 resilience to jamming attacks," in *Proceedings of the 26th IEEE International Conference on Computer Communications (INFOCOM'07)*, pp. 2526–2530, Barcelona, Spain, May 2007.

[24] J. Liu, Y. Xiao, S. Li, W. Liang, and C. L. P. Chen, "Cyber security and privacy issues in smart grids," *IEEE Communications Surveys and Tutorials*, vol. 14, no. 4, pp. 981–997, 2012.

[25] E. Cayirci and C. Rong, *Security in Wireless Ad Hoc and Sensor Networks*, John Wiley and Sons, Ltd, 2009.

[26] H. K. D. Sarma and A. Kar, *Security Threats in Wireless Sensor Networks*, Elsevier, October 2006.

[27] M. Tyndall, R. Marshall, E. K. Armstrong, and C. Marshman, "Potential EMC implementation problems of smart metering, display and communications," in *Proceedings of the 2011 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies, ISGT Europe 2011*, Manchester, UK, December 2011.

[28] R. Smolenski, *Conducted Electromagnetic Interference (EMI) in Smart Grids*, Springer, London, UK, 2012.

[29] A. V. Pramo, M. Abdul Azeem, and O. M. Prakash, "Detecting the sybil attack in wireless sensor network," *International Journal of Computers & Technology*, vol. 3, no. 1, 2012.

[30] S. Ji, T. Chen, and S. Zhong, "Wormhole attack detection algorithms in wireless network coding systems," *IEEE Transactions on Mobile Computing*, vol. 14, no. 3, pp. 660–674, 2015.

[31] A. D. Wood and J. A. Stankovic, "Denial of service in sensor networks," *Computer*, vol. 35, no. 10, pp. 54–62, 2002.

[32] K. Billewicz, *Smart Metering Inteligentny System Pomiarowy*, Wydawnictwo Naukowe PWN, 2011.

[33] A. Lee and T. Brewer, "Guidelines for smart grid cyber security, 1, smart grid cyber security strategy, architecture and high-level requirements," NISTIR 7628, 2010.

[34] M. Esposito, C. Mazzariello, F. Oliviero, S. P. Romano, and C. Sansone, "Evaluating pattern recognition techniques in intrusion detection systems," in *Proceedings of the 5th International Workshop on Pattern Recognition in Information Systems (PRIS'05), in Conjunction with ICEIS 2005*, pp. 144–153, May 2005.

[35] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: a survey," *ACM Computing Surveys*, vol. 41, no. 3, article 15, 2009.

[36] M. Xie, S. Han, B. Tian, and S. Parvin, "Anomaly detection in wireless sensor networks: a survey," *Journal of Network and Computer Applications*, vol. 34, no. 4, pp. 1302–1325, 2011.

[37] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences*, vol. 19, no. 1, pp. 15–18, 1977.

[38] E. S. Gardner Jr., "Exponential smoothing: the state of the art-part II," *International Journal of Forecasting*, vol. 22, no. 4, pp. 637–666, 2006.

[39] E. S. Gardner, "Exponential smoothing: The state of the art," *Journal of Forecasting*, vol. 4, no. 1, pp. 1–28, 1985.

[40] C. C. Pegels, "Exponential forecasting: some new variations," *Management Science*, vol. 12, pp. 311–315, 1969.

[41] O. L. Davies and R. G. Brown, "Statistical forecasting for inventory control," *Journal of the Royal Statistical Society. Series A (General)*, vol. 123, no. 3, p. 348, 1960.

[42] C. C. Holt, *Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages, ONR Memorandum*, vol. 52,

Carnegie Institute of Technology. Available from the Engineering Library, University of Texas at Austin, Pittsburgh, PA, USA, 1957.

[43] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science. Journal of the Institute of Management Science. Application and Theory Series*, vol. 6, pp. 324–342, 1960.

[44] R. G. Brown, *Smoothing, Forecasting and Prediction of Discrete Time Series*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1963.

[45] J. Bollinger, *Bollinger on Bollinger Bands*, McGraw Hill, 2002.

[46] S. Vervoort, "Smoothing the bollinger bands," *Technical Analysis of Stocks & Commodities*, vol. 28, no. 6, pp. 40–44, 2010.

[47] UTP University of Science and Technology in Bydgoszcz, Poland, http://wyszukaj.utp.edu.pl/mapa.

[48] IEC 61000-4-4, http://www.iec.ch/emc/basic_emc/basic_emc_immunity.htm.

[49] Y. Guo, C.-W. Ten, S. Hu, and W. W. Weaver, "Modeling distributed denial of service attack in advanced metering infrastructure," in *Proceedings of the 2015 IEEE Power and Energy Society Innovative Smart Grid Technologies Conference (ISGT'15)*, February 2015.

[50] P. Cheng and M. Zhu, "Lightweight anomaly detection for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 2015, Article ID 653232, 2015.

[51] V. Garcia-Font, C. Garrigues, and H. Rifà-Pous, "A comparative study of anomaly detection techniques for smart city wireless sensor networks," *Sensors (Switzerland)*, vol. 16, no. 6, article 868, 2016.