

Songtao Guo · Kai Liu ·
Chao Chen · Hongyu Huang (Eds.)

Communications in Computer and Information Science

1101

Wireless Sensor Networks

13th China Conference, CWSN 2019
Chongqing, China, October 12–14, 2019
Revised Selected Papers

 Springer



Communications in Computer and Information Science


1101

Commenced Publication in 2007

Founding and Former Series Editors:

Phoebe Chen, Alfredo Cuzzocrea, Xiaoyong Du, Orhun Kara, Ting Liu,
Krishna M. Sivalingam, Dominik Ślęzak, Takashi Washio, Xiaokang Yang,
and Junsong Yuan

Editorial Board Members

Simone Diniz Junqueira Barbosa 

*Pontifical Catholic University of Rio de Janeiro (PUC-Rio),
Rio de Janeiro, Brazil*

Joaquim Filipe 

Polytechnic Institute of Setúbal, Setúbal, Portugal

Ashish Ghosh

Indian Statistical Institute, Kolkata, India

Igor Kotenko 

*St. Petersburg Institute for Informatics and Automation of the Russian
Academy of Sciences, St. Petersburg, Russia*

Lizhu Zhou

Tsinghua University, Beijing, China

More information about this series at <http://www.springer.com/series/7899>

Songtao Guo · Kai Liu ·
Chao Chen · Hongyu Huang (Eds.)

Wireless Sensor Networks

13th China Conference, CWSN 2019
Chongqing, China, October 12–14, 2019
Revised Selected Papers

Editors

Songtao Guo
Chongqing University
Chongqing, China

Chao Chen
Chongqing University
Chongqing, China

Kai Liu
Chongqing University
Chongqing, China

Hongyu Huang
Chongqing University
Chongqing, China

ISSN 1865-0929

ISSN 1865-0937 (electronic)

Communications in Computer and Information Science

ISBN 978-981-15-1784-6

ISBN 978-981-15-1785-3 (eBook)

<https://doi.org/10.1007/978-981-15-1785-3>

© Springer Nature Singapore Pte Ltd. 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd. The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

The China Conference on Wireless Sensor Networks (CWSN) is the annual conference of CCF on Internet of Things (IoT). As a leading conference in the field of IoT, CWSN is the premier forum for IoT researchers and practitioners from academia, industry, and government in China to share their ideas, research results, and experiences, which highly promotes research and technical innovation in these fields domestically and internationally.

CWSN 2019 provided an academic exchange, research, and development forum for IoT researchers, developers, enterprises, and users to exchange ideas and experiences in IoT research and application. The conference also explored key challenges facing IoT research and application and research hotspots. As a high-level forum for the design, implementation, and application of IoT, CWSN 2019 aimed to promote the exchange and application of IoT theory and technology. Top experts at home and abroad presented special reports, and IoT-related companies showcased their latest technologies.

This year, CWSN received 158 submissions. After a thorough reviewing process, 31 English papers and 65 Chinese papers were selected for presentation as full papers. The high-quality program would not have been possible without the authors who chose CWSN 2019 as a venue for their publications. We are also very grateful to the members of the Program Committee and Organizing Committee, who put a tremendous amount of effort into soliciting and selecting research papers with a balance of high quality, new ideas, and new applications.

We hope that you enjoy reading and benefit from the proceedings of CWSN 2019.

October 2019

Songtao Guo

Organization

CWSN 2019 was organized by China Computer Federation, the Internet of Things Professional Committee of CCF and Chongqing University.

Organizing Committee

Conference Co-chairs

Jianzhong Li	China Computer Federation, China
Xiaofeng Liao	Chongqing University, China
Yuanyuan Yang	The State University of New York at Stony Brook America, USA

Conference Honorary Chair

Hao Dai	Academician of Chinese Academy of Engineering, China
---------	---

Program Co-chairs

Huadong Ma	Beijing University of Posts and Telecommunications, China
Songtao Guo	Chongqing University, China

Program Vice-chairs

Chao Chen	Chongqing University, China
Kai Liu	Chongqing University, China
Liang Feng	Chongqing University, China

Chairmen of the Excellent Paper Awards

Xue Wang	Tsinghua University, China
Li Cui	Institute of Computing Technology, Chinese Academy of Sciences, China

Chairmen of the Excellent Young Scholars Forum

Huadong Ma	Beijing University of Posts and Telecommunications, China
Limin Sun	Institute of Information Engineering, Chinese Academy of Sciences, China

Organizing Co-chairs

Tao Xiang Chongqing University, China
Kunyin Guo Chongqing University, China

Organizing Committee

Weijie Shen Chongqing University, China
Hongyu Huang Chongqing University, China
Chaochan Xiang Chinese People's Liberation Army Service Academy,
China
Guangchao Yang Chongqing University, China
Huiwei Wang Southwest University, China
Wang Ying Southwest University, China

Program Committee

Guangwei Bai Nanjing University of Technology, China
Ming Bao Institute of Acoustics, Chinese Academy of Sciences,
China
Qingsong Cai Beijing Technology and Business University, China
Shaobin Cai Huaqiao University, China
Bin Cao Harbin Institute of Technology and Shenzhen Graduate
School, China
Fanzai Zeng Hunan University, China
Guihai Chen Nanjing University, China
Hong Chen Renmin University of China, China
Jiaxing Chen Hebei Normal University, China
Xi Chen National Grid Information and Communication Branch,
China
Xiaojiang Chen Northwest University, School of Information, China
Yongle Chen Taiyuan University of Technology, China
Zhikui Chen Dalian University of Technology, China
Li Cui Chinese Academy of Sciences, China
Xunxue Cui Army Officer School, China
Zhidong Deng Tsinghua University, China
Wei Dong Zhejiang University, China
Hongwei Du Harbin Institute of Technology and Shenzhen Graduate
School, China
Dingyi Fang Northwest University, China
Xiufang Feng Taiyuan University of Technology, China
Deyun Gao Beijing Jiaotong University, China
Hong Gao Harbin Institute of Technology, China
Jibin Gong Yan Shan University, China
Songtao Guo Chongqing University, China
Zhongwen Guo China Ocean University, China
Guangjie Han Hohai University, China

Yanbo Han	Institute of Data Engineering and North China University of Technology, China
Daojing He	East China Normal University, China
Shibo He	Zhejiang University, China
Chengquan Hu	Jilin University, China
Yanjun Hu	Anhui University, China
Qiangsheng Hua	Huazhong University of Science and Technology, China
He Huang	Suzhou University, China
Liusheng Huang	China University of Science and Technology, China
Hongbo Jiang	Huazhong University of Science and Technology, China
Qi Jing	Peking University, School of Software and Microelectronics, China
Bo Jing	Air Force Engineering University, China
Deying Li	Renmin University of China, China
Fan Li	Beijing Institute of Technology, China
Fangmin Li	Wuhan University of Technology, China
Guanghui Li	Jiangnan University, China
Guorui Li	Northeastern University Qinhuangdao Branch, China
Hongwei Li	University of Electronic Science and Technology, China
Jianbo Li	Qingdao University, China
Jianzhong Li	Harbin Institute of Technology, China
Jinbao Li	Heilongjiang University, China
Minglu Li	Shanghai Jiaotong University, China
Renfa Li	Hunan University, China
Shining Li	Northwestern Polytechnical University, China
Xiangyang Li	University of Science and Technology of China, China
Zhetao Li	Xiangtan University, China
Liang Hongbin	Southwest Jiaotong University, China
Yongzhen Liang	Changzhou University, China
Wei Liang	Institute of Chinese Academy of Sciences Shenyang Institute of Automation, China
Yaping Lin	Hunan University, China
Jiajia Liu	Xi'an University of Electronic Science and Technology, China
Liang Liu	Beijing University of Posts and Telecommunications, China
Min Liu	Institute of Computing Technology, Chinese Academy of Sciences, China
Xingcheng Liu	Zhongshan University, China
Yunhao Liu	Tsinghua University, China
Xiang Liu	Peking University, China
Zhaohua Long	Chongqing University of Posts and Telecommunications, China

Juan Luo	Hunan University, China
Huadong Ma	Beijing University of Posts and Telecommunications, China
Li Ma	North Industrial University, China
Jianwei Niu	Beijing University of Aeronautics and Astronautics, China
Xiaoguang Niu	Wuhan University, China
Jian Peng	Sichuan University, China
Peng Li	Jiangnan University, China
Shaoliang Peng	National University of Defense Technology, China
Wangdong Qi	PLA University of Science and Technology, China
Fengyuan Ren	Tsinghua University, China
Shikai Shen	Kunming University, China
Yulong Shen	Xi'an University of Electronic Science and Technology, China
Jian Shu	Nanchang Aviation University, China
Lijuan Sun	Nanjing University of Posts and Telecommunications, China
Limin Sun	Institute of information Engineering, Chinese Academy of Sciences, China
Dan Tao	Beijing Jiaotong University, China
Liqin Tian	North China Institute of Science and Technology, China
Yang Wang	China University of Science and Technology, China
Kun Wang	Nanjing University of Posts and Telecommunications, China
Lei Wang	Dalian University of Technology, China
Liangmin Wang	Jiangsu University, China
Ping Wang	Chongqing University of Posts and Telecommunications, China
Chuanchuan Wang	Nanjing University of Posts and Telecommunications, China
Xiaoming Wang	Shaanxi Normal University, China
Xiaodong Wang	National University of Defense Technology, China
Xinbing Wang	Shanghai Jiaotong University, China
Xue Wang	Tsinghua University, China
Yiding Wang	North China University of Technology, China
Yuexuan Wang	Tsinghua University, China
Zhibo Wang	Wuhan University, China
Zhi Wang	Zhejiang University, China
Zhu Wang	Harbin Institute of Technology, China
Wei Wei	Xi'an University of Technology, China
Xingjun Wu	Tsinghua Tongfang Microelectronics, China
Xiaojun Wu	Shaanxi Normal University, China
Deqin Xiao	South China Agricultural University, China
Fu Xiao	Nanjing University of Posts and Telecommunications, China

Kun Xie	Hunan University, China
Yongping Xiong	Beijing University of Posts and Telecommunications, China
Guangtao Xue	Shanghai Jiaotong University, China
Geng Yang	Nanjing University of Posts and Telecommunications, China
Weidong Yang	Henan University of Technology, China
Weidong Yi	University of Chinese Academy of Sciences, China
Ruiyun Yu	Northeast University, China
Jiguo Yu	Qufu Normal University, China
Shigeng Zhang	Central South University, China
Shuqin Zhang	Zhongyuan Institute of Technology, China
Yunzhou Zhang	Northeastern University, China
Junhui Zhao	East China Jiaotong University, China
Zenghua Zhao	Tianjin University, China
Jiping Zheng	Nanjing University of Aeronautics and Astronautics, China
Hongzi Zhu	Shanghai Jiaotong University, China
Hongsong Zhu	Chinese Academy of Sciences, China
Yihua Zhu	Zhejiang University of Technology, China
Liehuang Zhu	Beijing Institute of Technology, China
Shihong Zou	Beijing University of Posts and Telecommunications, China
Wei Chen	Beijing Jiaotong University, China
Haiming Chen	Ningbo University, China
Honglong Chen	China University of Petroleum (East China), China
Xu Chen	Zhongshan University, China
Silao Cheng	Harbin Institute of Technology, China
Kaikai Chi	Zhejiang University of Technology, China
Xiaochao Dang	Northwest Normal University, China
Guangsheng Feng	Harbin Engineering University, China
Zhitao Guan	North China Electric Power University, China
Zhanjun Hao	Northwest Normal University, China
Jie Jia	Northeastern University, China
Feng Li	Shandong University, China
Jie Li	Northeastern University, China
Yanjun Li	Zhejiang University of Technology, China
Zhuo Li	Beijing Information Science and Technology University, China
Tie Qiu	Dalian University of Science and Technology, China
Yiran Shen	Harbin Engineering University, China
Xiaoxia Song	Shanxi Datong University, China
Xiaohua Tian	Shanghai Jiaotong University, China
Qingshan Wang	Hefei University of Technology School of Mathematics, China
Tian Wang	Huaqiao University, China

Hejun Wu	Zhongshan University, China
Ling Xiao	Hunan University, China
Lei Xie	Nanjing University, China
Yuan Yan	China Internet of Things Research and Development Center, China
Guisong Yang	University of Shanghai for Science and Technology, China
Zuwei Yin	PLA Information Engineering University, China
Ju Zhang	PLA (People's Liberation Army) of China, Branch 61785, China
Lei Zhang	Tianjin University, China
Lichen Zhang	Shaanxi Normal University, China
Lianming	Zhang Hunan Normal University, China
Jumin Zhao	Taiyuan University of Technology, China
Anfu Zhou	Beijing University of Posts and Telecommunications, China
Changbing Zhou	China University of Geosciences, China

Organizers

Organized by



China Computer Federation, China

Co-organized by

The Internet of Things Professional Committee of CCF

Hosted by



重慶大學
CHONGQING UNIVERSITY

Chongqing University

Sponsoring Institutions



Huawei Technologies Co., Ltd.



Chongqing Xiewen Science Co., Ltd.



Beijing Qi'anxin Technology Co., Ltd.



Inspur Group



Chongqing University of Posts
and Telecommunications



Wuxi Tsinghua Internet of Things Center



Zhongzhixun (Wuhan) Technology Co., Ltd.



Chongqing Bingrun Technology Co., Ltd.

Contents

Fundamentals on Internet of Things

Improving the Scalability of LoRa Networks Through Dynamical Parameter Set Selection	3
<i>Qingsong Cai and Jia Lin</i>	
A Weighted Voronoi Diagram Based Self-deployment Algorithm for Heterogeneous Mobile Sensor Network in Three-Dimensional Space	19
<i>Li Tan, Xiaojiang Tang, Minghua Yang, and Haoyu Wang</i>	
Joint Uplink and Downlink Optimization for Resource Allocation Under D2D Communication Networks.	35
<i>Di He, Guangsheng Feng, Bingyang Li, Hongwu Lv, Huiqiang Wang, and Quanming Li</i>	
FaLQE: Fluctuation Adaptive Link Quality Estimator for Wireless Sensor Networks.	48
<i>Wei Liu, Yu Xia, and Rong Luo</i>	
Rate Adaptive Broadcast in Internet of Things	61
<i>Linghe Kong, Zhe Wang, Yongshuai Duan, Tong Meng, Fan Wu, and Guihai Chen</i>	
An IOT Data Collection Mechanism Based on Cloud-Edge Coordinated Deep Learning	76
<i>Zi-hao Wang and Jing Wang</i>	
A Malicious Anchor Detection Algorithm Based on Isolation Forest and Sequential Probability Ratio Testing (SPRT).	90
<i>Jun Peng and Xingcheng Liu</i>	
Noisy Data Gathering in Wireless Sensor Networks via Compressed Sensing and Cross Validation	101
<i>Xiaoxia Song, Yong Li, and Wenmei Nie</i>	
Fuzzy-K: Energy Efficient Fuzzy Clustering Routing Protocol Based on Cross-Technology Communication in Wireless Sensor Network	112
<i>Yue Yu, Fanrong Meng, and Ming Li</i>	
An Improved Method of Pending Interest Table in Named Data Networking.	127
<i>Peiyuan Gu, Yabin Xu, and Tian Song</i>	

Applications on Internet of Things

Real-Time Bridge Structural Condition Evaluation Based on Data Compression. 143
Jingpei Dan, Ling Liu, Yuming Wang, Junji Chen, and Xia Huang

Task Assignment Algorithm Based on Social Influence in Mobile Crowd Sensing System 154
Anqi Lu and Jinghua Zhu

A Barrier Coverage Enhancement Algorithm in 3D Environment 169
Xiaochao Dang, Yuexia Li, Zhanjun Hao, and Tong Zhang

Sensor-Cloud Based Precision Sprinkler Irrigation Management System. 184
Mingzheng Zhang, Shuming Xiong, and Liangmin Wang

Deep Memory Network with Auxiliary Sequences for Chinese Implied Sentiment Analysis 198
Chao Wang, Yunhua He, Limin Sun, Chengjie Pang, and Jitong Li

Intelligent Traffic Light System for High Priority Vehicles. 212
Guiduan Li, Guozhi Song, and Wen Li

Personalized Recommendation Based on Tag Semantics in the Heterogeneous Information Network. 224
Bin Yan, Lichen Zhang, Longjiang Guo, Meirei Ren, and Ana Wang

High-Quality Learning Resource Dissemination Based on Opportunistic Networks in Campus Collaborative Learning Context 236
Peng Li, Hong Liu, Longjiang Guo, Lichen Zhang, Xiaoming Wang, and Xiaojun Wu

IntelliSense, Location and Tracking

Integrated Redundant APs Reduction and Transfer Learning for Indoor WLAN Intrusion Detection via Link-Layer Data Transformation. 251
Xinyue Li, Mu Zhou, Yaoping Li, Hui Yuan, and Zengshan Tian

A Near-Optimal Heterogeneous Task Allocation Scheme for Mobile Crowdsensing 263
Guangsheng Feng, Quanming Li, Junyu Lin, Hongwu Lv, Huiqiang Wang, and Silin Lv

A Lightweight Neural Network Localization Algorithm for Structureless Wireless Sensor Networks 275
Rong Gao, Zhongheng Yang, and Hejun Wu

A Fast Offline Database Construction Mechanism for Wi-Fi Fingerprint Based Localization Using Ultra-Wideband Technology 289
Huilin Jie, Hao Zhang, Kai Liu, Feiyu Jin, Chao Chen, and Chaocan Xiang

A Moving Target Trajectory Tracking Method Based on CSI 302
Zhanjun Hao, Lihua Yan, and Xiaochao Dang

A CSI-Based Indoor Intrusion Detection and Localization Method 317
Xiaochao Dang, Caixia Li, Zhanjun Hao, and Yuan Cao

Wi-SD: A Human Motion Recognition Method Based on CSI Amplitude and Phase Information 332
Xiaochao Dang, Tong Zhang, Zhanjun Hao, and Yuexia Li

Data-Quality-Aware Participant Selection Mechanism for Mobile Crowdsensing 348
Hongbin Sun and Dan Tao

Infrared Small Target Detection Based on Facet-Kernel Filtering Local Contrast Measure 360
Peng Du and Askar Hamdulla

Author Index 369

Fundamentals on Internet of Things



Improving the Scalability of LoRa Networks Through Dynamical Parameter Set Selection

Qingsong Cai and Jia Lin^(✉)

School of Computer and Information Engineering,
Beijing Technology and Business University, Beijing 100048, China
lemonlinjia@126.com

Abstract. LoRa technology has emerged as an interesting solution for Low Power Wide Area applications. To support a massive amount of devices in large-scale networks, it is necessary to design an appropriate parameter allocation scheme for device. LoRa devices provide high flexibility in choosing settings of communication parameters (including spreading factors, bandwidth, coding rate, transmission power, etc), which results in there are over 6000 settings for choosing. However, the existing methods mainly focus on the same parameter setting for network deployment. To this aim, the impact of different parameter selections on communication performance is analyzed first. Then, channel collision and link budget model are established and implemented in the NS3 simulator. A dynamic parameter selection method based on orthogonal genetic algorithm (OGA) is introduced to solve the model, ultimately according to link budget, each device selects its parameter setting, which minimized collision probability. Finally, simulation results show that the OGA algorithm proposed in this paper can improve the packet delivery rate by 30%. Knowing different packet sizes have an impact on network performance, the experiment also evaluated the impact of different packet sizes on network transmission reliability under different parameter setting methods, the introduced OGA has significantly improved adaptability and scalability of the network in the case of high payloads.

Keywords: Internet of things · LoRa · Low power wide-area network · Orthogonal genetic algorithm · Parameter combination

1 Introduction

With the continuous development of IoT, its application domains are increasing, and the number of device deployment is exploding. According to forecasts, the number of connected IoT devices will continue to grow at a rate of 32% per year and it is estimated that there will be 500 billion devices connected with wireless communication by 2022 [1]. Compared with the Internet, some emerging IoT applications require merely less memory, bandwidth and processing ability of the

device to efficiently complete their work [2,3]. Traditional cellular networks and related technologies have been unable to satisfy the demands for “less of everything” in terms of network capacity, communication range, energy consumption and cost. Aiming at solving this issue, LPWAN, a very attractive and promising communication technology, is utilized to carry out the long-range communication with a large number of devices. And it is gradually applied on the smart home, metering application and other application fields [4-6]. LPWAN has the advantage on offering long-range connectivity for low power and low rate end devices. Therefore, it is suitable for those applications that are delay-tolerant, only need low data rates and typically require low power consumption. Recently, small payload packets with low amount of data have been garnered widespread attention in the IoT industry [7]. For example, for smart meters, each device is only required to transmit a packet per day. Other similar applications merely need transmit small amounts of discrete packet such as temperature and humidity. Such applications regularly use low-cost and low-consumption processors. When using LPWAN technology to transmit packets, the device can operate for several years with only one battery.

As an emerging LPWAN technology, LoRa has attracted much attention not only for its low power consumption and low deployment cost, but also for transfer of messages over a long-range. LoRa is a physical layer technique. Its MAC layer solution is regarded as LoRaWAN [8]. To keep the system free of complex routing protocols, LoRa technologies often rely on star topologies, in which end devices communicate directly with gateways in a single hop [9]. In order to achieve the scalability of LoRa network, LoRa technologies are required to provide connectivity for a massive number of IoT devices. And a large number of parameter settings will be generated from massive IoT devices which need to meet different IoT applications with varying communication patterns. LoRa provides a range of communication parameter settings, including Spreading Factor (SF), Bandwidth (BW), Code Rate (CR), Transmission Power (TP) and Packet Size (PS) [10]. Many combination settings are orthogonal and keep communications from simultaneous collision, it has to be noted that using the same parameters increases the probability of collision. A packet can have significant variations in ToA (Time on Air) depending on the selected setting. For example, a 20 bytes packet can vary between 7 ms and 2.2 s. For this reasons it is indispensable in a LoRa network that end devices with battery-powered make good transmission parameter choices.

In view of the above problems, some researches have proposed SF adaptive optimization strategy [8,11,12], which can effectively improve the fairness of Packet Error Rate (PER) in LoRa network. However, this method does not take into account the full spectrum of parameters governing such as BW, CR, TP and PS on scalability of LoRa networks. Considering the shortcomings of existing methods, in this paper, a model of channel collision and link budget for single-gateway LoRa network is established and implemented in NS3 simulator. By solving the model, the parameter combination with minimum collision probability obtained. Algorithm for solving multi-objective combinatorial problems with

exhaustion method, genetic algorithm, simulated annealing algorithm. Considering that the end device is battery powered, the exhaustive method consumes too long time and is not suitable for this scenario. Simulated annealing algorithm is a random algorithm with slow convergence speed and cannot find the global optimal solution. The Genetic Algorithm (GA) processes the individual chromosomes encoded by the parameter set and can simultaneously process multiple individuals in the population. However, the core operation-crossover operator of the traditional GA is random in factor segmentation. The search has a certain blindness to some extent, which reduces the search efficiency of the algorithm. Therefore, the orthogonal design method is utilized to design crossover operator, and as a result, crossover operator self-adaptive to adjust the location for dividing the parents into several sub-vectors, so as to generate the population of genetic algorithm. In order to achieve a higher user experience quality at a lower power consumption, by using the dynamic parameter selection method based on orthogonal genetic algorithm (OGA), the communication device can independently select SF, BW, CR, TP and other parameters. In summary, the main work of this paper includes:

- First of all by experiment, a study of the impact of the LoRa transmission parameters SF, BW, CR, TP and PS on communication performance is analyzed;
- A channel collision and link budget model are established for single-gateway LoRa network and implemented in NS3 simulator. An OGA-based dynamic parameter selection algorithm is proposed. By using the cross-combination dynamic selection of parameters, the collision probability is minimized;
- By simulating 10,000 end devices with the same PS, PDR was 30% better than the static deployment. Knowing different packet sizes have an impact on network performance, the experiment evaluated the impact of different PS on network transmission reliability under different parameter setting methods.

2 Related Work

To optimize the performance of LoRa, many work mainly focus on how to allocate the wireless resources effectively. Author Peng et al. [11] at SIGCOMM 2018 proposed PLoRa, an ambient backscatter design that enables long-range wireless connectivity for battery less IoT devices, but the author points out in the study that the limitation of PLoRa design is that only encoding a data rate (determined by SF, BW and CR). There are also many other works studied on performance improvements of network-related parameters. To sum up, the research on LoRa parameter configuration can be divided into two categories:

The first category is the static deployment method which refers to a setting where all end devices employ the same parameter. Martin et al. [9] adopted three static parameter settings for network communication, those are SN1, SN2, SN3. SN1, which is the longest ToA setting, SN2 is the shortest ToA setting, and SN3 is the default setting. These parameter settings are applied to evaluate the scalability of the LoRa network for the established link model. Thiemo et al. solved

the problem of interference caused by deploying multiple independent LoRa networks in close proximity [12]. All experiments have the same parameter setting SN1 and SN3. SN3 is similar to SN1 except for a lower CR which reduces the ToA and leads to fewer collisions. A stronger CR is very energy-costly as packets contain redundant information, however, better communication performance can be achieved in areas with burst interference. It is possible to conclude that with the increase number of end devices, static parameter deployment cannot make the LoRa scaled well. Therefore, the dynamic configuration and autonomous selection of LoRa parameters are of great significance for reducing energy consumption of end device and improving network scalability.

The second category is the dynamic deployment method. Martin et al. [9] evaluated the impact of dynamic communication parameter selection on PDR. Three settings SN3, SN4, and SN5 are compared. SN3 is the same as the experiment in static deployment as a comparative experiment. In the case of SN4, set BW, SF, and CR to minimize ToA (with a constant TP = 14 dBm), SN5 sets the first ToA determined by BW, SF, CR and minimizes the selection of the TP. Dynamic parameter settings have significant improvements over static setting implementation in LoRa. But it has to be considered that this achievement is not practical and relies on quite optimistic assumptions. First, the minimum ToA setting has the lowest CR, which fail to provide sufficient protection. Second, due to environmental changes, it is necessary to re-evaluate selected settings from time to time and requires implementation of complex protocols to facilitate settings in the LoRa network. EXPLoRa heuristic method [13] aims to effectively allocate SFs between end devices, and proposes two SF allocation methods, EXP-SF and EXP-AT. EXP-SF distributed SFs equally among N nodes based on RSSI, EXPLoRa-AT is a more sophisticated method of transmitting data packets across SF channels by equalized ToAs. The two aforementioned methods use same BW, CR and TP, which leads to a higher overall data rate than in reality. Martin et al. [14] developed a link probing regime which enables us to quickly determine transmission parameter selection with lower energy consumption, but they did not evaluate the impact of different parameters on network scalability. Adelantado et al. [15] reported the characteristics and limits of LoRa according to the relationship between duty cycle and throughput of different PS. Taoufik et al. [16] pointed out that the consumed energy changes with different LoRa parameters such as SF, CR, BW, TP and PS. Optimizing these parameters are of great importance for both reducing sensor energy consumption and network scalability.

Although there have been studies on SF dynamic allocation, this category is still classified as static deployment in our study because other parameters such as BW, CR and TP are indeed fixed.

3 Overview of LoRa

3.1 LoRa Physical Layer Parameter

LoRa uses the CSS modulation combined with the Forward Error Correction (FEC) to maintain low power and long communication range [17]. LoRa key properties are: long-range, high robustness, multipath resistance, Doppler resistance and low power [18]. A typical LoRa device provides several main modulation parameters SF, BW, CR and TP. These parameters influence the effective bit-rate of the modulation, its resistance to interference noise and ease of decoding. LoRa's communication performance can be tuned by varying the selection of these parameter settings and the LoRa network scalability is determined by this key factors. BW is the width of frequencies in the transmission band. The higher is the BW, the shorter is the ToA and the lower is the sensitivity. SF is the number of bits encoded into each symbol, namely the ratio of chip rate to symbol rate. Each increase in SF allows a longer communication range, but doubles ToA and ultimately energy consumption. CR is the FEC code rate used by LoRa modem to protect from a burst of interference. Depending on the CR selected, additional robustness can be obtained with interference. A higher CR offers more protection, but increases ToA. A high TP will result in a higher RSSI, increasing the range of communication while allowing a lower PER. Maximum PS for a LoRa network is 255 bytes. In our testing, the packet sizes were configured as: 10 bytes, 20 bytes, 30 bytes, 40 bytes, 60 bytes, 80 bytes.

3.2 The Effect of Different Parameter

Each data rate, through the combination of SF, BW, CR and PS, experiences different ToA, thus different collision probability. Following these considerations, we analyzed the effects of different parameters on ToA and energy consumption through experiments. Figure 1 shows the comparison between ToA and energy consumption with different parameters. It can be seen from the figure that the ToA of data packets varies significantly with different communication parameter selections. For example, the ToA of packets of 20 bytes with different parameters varies between 7 ms–2 s. Therefore, the dynamic selection of communication parameters has a significant impact on the scalability of LoRa network deployment.

4 System Modeling

In order to study the influence of communication parameters on LoRa network performance, in this section, we describe system model that are used in the paper.

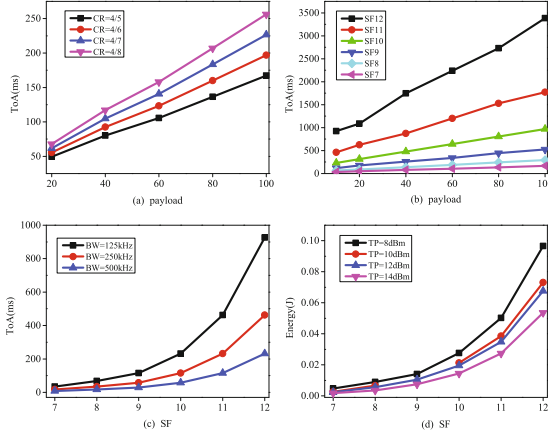


Fig. 1. Effect of different parameter on ToA and Energy.

4.1 Channel Collision Model

The LoRaWAN physical layer supports adjustable SFs, $SFs \in [7, 12]$ and spread spectrum signals with different SFs have good orthogonality and can transmit in the same channel simultaneously without interference [19]. As shown in Fig. 2, in LoRaWAN, it is assumed that the available SFs are SF9 and SF11, and the following three cases are transmitted in the channel. In case 1, on account of two SFs 11 arrive at different channels simultaneously, they can be successfully received and decoded. In the case 2, SF9 and SF11 are in the same channel, and due to their orthogonality, collisions can be avoided. In the case 3, the same SF9 arrives at the same channel simultaneously, causing a collision to occur. When there are multiple SFs transmitting on the channel, two collide occur: Two packets with same SF arriving at the same channel simultaneous cause a collision, thus causing data packet loss. Collisions with the same SF: the probability of at least one collision with the same SF using the random access formula, as shown in Eq. (1):

$$P_{coll,sf} = 1 - e^{-2G_{sf}} \quad (1)$$

where G_{sf} is the amount of packets generated during the transmission of one packet with SF.

The transmission time of a packet T_{sf} in LoRa is given by Eq. (2)

$$T_{sf} = \frac{L}{R_b} \quad (2)$$

R_b is bit-rate that can be expressed as:

$$R_b = SF \times \frac{BW}{2^{SF}} \times CR \quad (3)$$

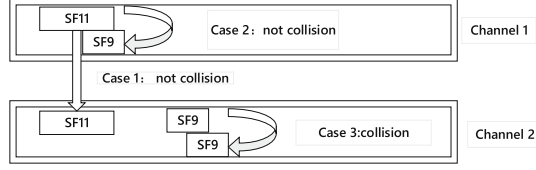


Fig. 2. Multiple access under the orthogonal SF.

the amount of traffic generated per unit of time λ is given by

$$\lambda = \frac{N}{T_i} \quad (4)$$

T_i the average packet inter arrival time per end device. The amount of traffic G_{sf} generated during the transmission of one packet using SF is

$$G_{sf} = \lambda \partial_{sf} T_{sf} \quad (5)$$

where ∂_{sf} is the fraction of devices using the same SF. Finally, the probability of collision can be expressed as

$$P_{coll,sf} = 1 - e^{-2[\frac{2^{sf}}{sf} \frac{L}{BW \times cr} \partial_{sf} \lambda]} \quad (6)$$

The probability of packet collision during transmission is closely related to the combination of these parameters and ∂_{sf} , which is the variable used to optimize the PDR. Therefore, the channel collision model is established as the objective function to improve PDR by dynamically selecting the combination of parameters, so as to improve the reliable transmission of the network.

4.2 Link Budget Model

The solution to our problem is not merely to minimize the probability of collision, but also to increase the correct reception of the gateway. The gateway is determining the reception depending on the receive sensitivity. These parameters also have significant influence on the receive sensitivity [20]. The increase of BW will reduce the sensitivity of the receiver, while the increase of SF will increase the sensitivity of the receiver. The evaluation and analysis of link budget depend on link parameters such as SNR and receive sensitivity, and receive sensitivity is positively correlated with SNR. The SNR is calculated as follows:

$$SNR = \frac{2^{sf} \times P_{rx}}{BW \times NF \times K \times T} \quad (7)$$

where P_{rx} , NF , K and T are the received power, receiver architecture noise figure, the Kelvin constant, the temperature respectively. The sensitivity can be defined in the following equation:

$$S_R(SF, BW) = -174 + 10\log_{10}(BW) + NF + SNR \quad (8)$$

Equation (8) depicts the L_{path} expression, Thus, real path loss in LoRa can be mathematically defined as:

$$L_{path} = \frac{P_{tx}}{S_R(SF, BW)} \quad (9)$$

In wireless communication, the MCL is defined as the maximum link budget allowed for each value of SF, BW and TP. To ensure correct signal demodulation, L_{path} must be smaller than MCL :

$$p = \begin{cases} 1 & MCL > L_{path} \\ 0 & else \end{cases} \quad (10)$$

where MCL is determined based on the sensitivity of the gateway and TP used by the end device as follows:

$$MCL = TP - S_R(SF, BW) \quad (11)$$

4.3 Packet Delivery Rate

In this paper, we define reliability as the ratio of the packets successfully received by the gateway to the total number of packets transmitted from the end device over a period of time. The main evaluation metric used in the simulation to estimate the performance is called packet delivery rate (PDR). The packets transmitted from the end device fail to deliver if they could not satisfy the condition mentioned in (10). The L_{path} defined in (9) refers to the average statistical path loss in the city environment. However, in real-life scenarios, path loss fluctuates and exact path loss is hard to estimate. This happens because wireless communication is effected by several unpredictable factors [21–23] such as the blocking of signals caused by large obstacles, fluctuation caused by weather conditions. Therefore, in order to simulate the real scenario as much as possible, we need to fully consider the impact of these parameters on communication.

5 Proposed Solution

In this section, we will explain in detail the proposed method. First, we specify the feasibility of the proposed method. Following it, we describe the implementation process of the algorithm. With small data rate and resource limited LoRa end device, configuration of optimal settings becomes challenging. Especially when the number of end device is large, how to utilize an appropriate dynamic selection algorithm becomes particularly important. For the LoRa parameter, the set $S = \{SF, BW, CR, TP\}$, where SFs $\in [7, 12]$ has 6 levels, BW = 125 kHz, 250 kHz and 500 kHz have 3 levels, CR = 4/5, 4/6, 4/7, 4/8 have 4 levels, and there are 13 levels in TP $\in [2, 14]$. When N nodes are considered, the search space for optimal parameter for each node is $N^{6 \times 4 \times 3 \times 13}$, this optimization problem can easily be solved by GA. The advantage of GA is reflected in the extensiveness

of the representation of the feasible solution. The object it deals with is not the parameter itself, but the individual gene obtained by encoding the parameter set and can simultaneously process multiple individuals in the population, which is very suitable for parameter configuration in LoRa network.

5.1 OGA Dynamic Selection

Although GA has been successfully applied in many optimization problems. However, a large number of studies have shown that traditional GA has many shortcomings [25], such as premature convergence and poor local search ability. Crossover operator imitates the natural process of chromosome gene recombination, and the core operation of GA. In the traditional way to carry out crossover operation, the location of factor segmentation is randomly generated and the search has certain blindness, which greatly reduces the search efficiency of the algorithm. In order to solve this problem, this paper adopts OGA as a solution to the problem, by integrating the orthogonal design into crossover operator, the position of the segmentation of the individual factors of the parent generation can be adjusted adaptively to generate the population of the GA. Among many combination parameters of LoRa, the two parent individuals involved in the crossover operation are: $p_1: (7, 125, 1, 2)$, $p_2: (8, 250, 2, 5)$. If the first and second dimensions of p_1 and p_2 are denoted as factor 1, third and fourth are considered as factor 2 respectively, the position of factor segmentation is shown in the dotted line in the equation. In this way, the cross operation of p_1 and p_2 is transformed into a two-factor, two-level experimental problem. Finally, orthogonal design is arranged in orthogonal table to generate offspring population p^1 , as shown

$$\begin{cases} p_1 = (7, |125, 1, 2) \\ p_2 = (8, |250, 2, 5) \end{cases} \Rightarrow p^1 = \begin{cases} (7, 125, 1, 2) \\ (7, 250, 2, 5) \\ (8, 125, 1, 2) \\ (8, 250, 2, 5) \end{cases} \quad (12)$$

After the three methods are utilized to segment the factors, eight descendant individuals including two fathers are generated. In the multi-point crossover operation, the cross combination method exists in various ways. As the number of intersections increases, the number of combination methods will increase sharply, and the number of intersections and the position of the cross operation are adaptively adjusted hence improve search efficiency. The specific method of factor segmentation is shown in step 1 of the algorithm. In the N-dimensional space, set Q parents involved in the recombination be p_1, p_2, \dots, p_Q . Each parent involved in the recombination was regarded as a level of the orthogonal design, namely the Q level. Then, each parent was divided into T groups, and each group was considered as one factor. In this way, the recombination problem of Q parent individuals is transformed into the orthogonal design problem of Q level and T factor. The whole algorithm flow is as follows: (1) The Q parent individuals involved in the recombination were considered as a level of the orthogonal design, and the i -th level was denoted as β_i , $i \in \{1, 2, 3, \dots, Q\}$ (2) The specific

grouping method is: randomly generate $T - 1$ integers, that is k_1, k_2, \dots, k_{T-1} , and satisfy the requirement that $1 < k_1 < k_2 < \dots < k_{T-1} < N$. Individual $X = (X_1, X_2, \dots, X_N)$ is divided into T parts, each of which represents a factor of individual X .

$$\begin{cases} f_1 = (x_1, x_2, \dots, x_{k_1}) \\ f_2 = (x_{k_1+1}, x_{k_1+2}, \dots, x_{k_2}) \\ \dots \\ f_T = (x_{k_{T-1}}, x_{k_{T-1}+1}, \dots, x_N) \end{cases} \quad (13)$$

Therefore, the Q levels of the i -th factor can be expressed as

$$\begin{cases} f_i(1) = (\beta_{k_{i-1}+1,1}, \beta_{k_{i-1}+2,1}, \dots, \beta_{k,1}) \\ f_i(2) = (\beta_{k_{i-1}+1,2}, \beta_{k_{i-1}+2,2}, \dots, \beta_{k,2}) \\ \dots \\ f_i(Q) = (\beta_{k_{i-1}+1,Q}, \beta_{k_{i-1}+2,Q}, \dots, \beta_{k,Q}) \end{cases} \quad (14)$$

(3) Select orthogonal table and M descendants were generated according to the orthogonal table

$$\begin{cases} (f_1(b_{1,1}), f_2(b_{1,2}), \dots, f_T(b_{1,T})) \\ (f_1(b_{2,1}), f_2(b_{2,2}), \dots, f_T(b_{2,T})) \\ \dots \\ (f_1(b_{M,1}), f_2(b_{M,2}), \dots, f_T(b_{M,T})) \end{cases} \quad (15)$$

Algorithm 1. Dynamic selection algorithm

```

1: function OPTIMAL SETTING( $P_{coll, sf}$ )
2:   Connection  $\leftarrow$  FALSE
3:    $S_i, T = 0$  initializes
4:   Select and crossover in (13)(14)(15)
5:   while ( $doT > 30$ )
6:      $S_i = \{SF, BW, CR, TP\}$ 
7:     Calculate  $Lpath$  given  $P_{coll, sf}$  from (9)
8:     while  $s$  in  $S_i$  do
9:       Calculate MCLs from (11)
10:      if  $MCLs > Lpath$  then
11:        Connection  $\leftarrow$  TRUE
12:        if  $P_{coll, sf} < 4\%$  then
13:           $S_i\{opt\} \leftarrow s$ 
14:        else
15:          return NULL
16:   return  $S_i$ 

```

The proposed OGA uses dynamic selection of optimal setting S_i based on estimated path loss, depending on the parameter combination with a minimum probability of collision, as per (11). In this paper, each settings is a

combination of independently varying SF, BW, CR and TP and thus is a subset of $S = \{\text{SF}, \text{BW}, \text{CR}, \text{TP}\}$. At the beginning, the proposed method initializes S_i by using orthogonal design method to generate the initial population $S = (7, 125, 1, 2), (7, 500, 4, 11), (8, 250, 2, 5), (8, 500, 3, 14), (9, 500, 3, 8), (9, 125, 2, 8), (10, 125, 4, 11), (10, 250, 1, 5), (11, 125, 3, 2), (11, 250, 4, 14)$, as it generate among all possible settings in LoRa transmission. Then set the maximum iterations $T = 30$, the initial iteration $T = 0$. After generating the population, it is first determined whether the number of iterations is reached. Random select two parent individuals to perform three cross operations at different positions. Calculate the fitness value (L_{path}) of the candidate solution, that is, the process of the above multi-point cross, hence generate the child generation population $t + 1$. The probability of selection = individual fitness value/total fitness value. The method then iterates over all possible setting and chooses the S_i which minimum collision possibility (fitness value) is within 4% for successful communication.

6 Evaluation and Results of Simulation Experiments

We used NS3 simulator to evaluate the scalability of LoRa networks. Since large-scale network deployment would be prohibitively expensive, it is not feasible to evaluate the scalability of such LoRa networks in real scenarios. For scalability, we focus on the capacity of a single gateway. This section presents the numerical results of the proposed method. We calculated PDR with respect to the total number of devices ranging from 0 to 10,000. The parameters in the experimental simulation are shown in Table 1.

Table 1. Parameters set.

Parameters	Values
End device	0–10000
<i>SF</i>	7–11
<i>BW</i>	125 kHz, 250 kHz, 500 kHz
<i>CR</i>	4/5, 4/6, 4/7, 4/8
λ	60 ms
<i>PL</i>	20B

6.1 The Impact of Parameter Selection

In this subsection, we first present the results of simulations with different SFs as a function of the number of end devices (up to 5000). Figure 3 shows PDR with different SFs from 7 to 12. We can see that PDR decreases when the

number of end device increases. The SF7 has better PDR compared with the others due to its short ToA but its range is reduced. Figure 4 shows the PDR with different SFs and CR. We can conclude that the higher CR (4/8) experience better PDR because it provide more protect. This means selection of parameters has a significant impact on network performance.

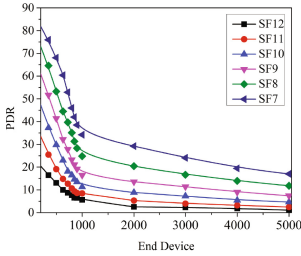


Fig. 3. The impact of SF selection.

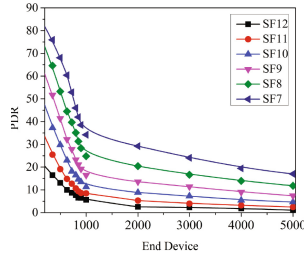


Fig. 4. The impact of CR and SF selection.

6.2 End Device Distribution Assessment

We define the ratio of devices using SF i as ∂_i . The motivation of our study is to optimize the PDR in environments where a large number of devices can use the same SF, thus the biased deployment must be considered. The optimized distribution of end devices ($\partial_7, \partial_8, \partial_9, \partial_{10}, \partial_{11}, \partial_{12}$) is (0.3, 0.1, 0.1, 0.2, 0.3, 0). The sum of the devices allocated by all SF is the total number of devices in the network. Although SF12 provides a larger coverage, it also increases the probability of collision, so we neglect its allocation. The network is serviced by a gateway located in the center of a circle with a radius of 8 km. The device distribution is shown in the Fig. 5.

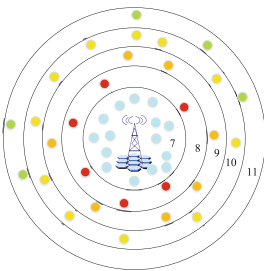


Fig. 5. End device distribution without SF = 12.

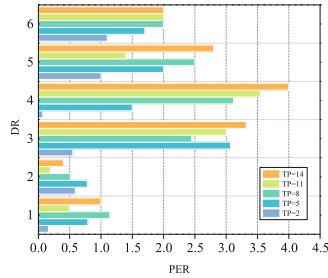


Fig. 6. PER control parameter selection within 4%.

6.3 Scalability Analyses

Before assessing PDR, first analyzes the PER of TP and DR determined by different SF, BW and CR, and the Fig. 6 of the DR value of 1 to 6. The SF, BW, and CR corresponding to the DR are shown in Table 2, according to the basic trend is rising with the increase of power, in the case of DR for 1 and 2, may be affected by the power of saturated, the overall PER are relatively low. Figure 6 shows the parameter selection with the set fitness value controlled within 4%.

Table 2. SF, BW and CR corresponding to different DRs.

DR	SF	BW	CR
1	11	125 kHz	4/6
2	10	250 kHz	4/5
3	9	125 kHz	4/8
4	8	500 kHz	4/5
5	7	125 kHz	4/7
6	7	250 kHz	4/5

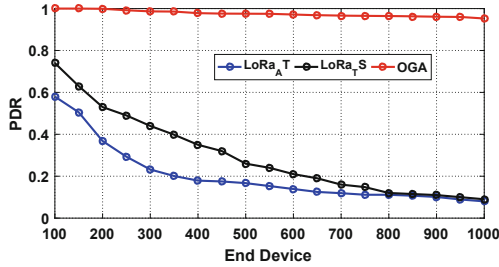


Fig. 7. Compare with dynamic parameter settings of PDR under 1000 device.

Figure 7 shows the PDR of the dynamic parameter method LoRa-TS and LoRa-AT under 1000 end device. With the increase number of end devices, the proposed method of dynamic parameter selection has a significant improvement on PDR. The Fig. 8 shows a comparison of PDRs for different dynamic methods of up to 10,000 end devices. The method in NS3, for each end device, by dynamically assigning the SF of lowest PER below a certain threshold and employed same parameters such as BW, CR and TP, the PDR based on OGA method was always high regardless of the number of end devices. In SN5, dynamic parameter selection is used, the minimum ToA and minimum TP are selected each time. When the number of end devices is less than 2000, there is no obvious difference

in PDR, but when the number of end devices is large, the OGA method proposed in this paper has obvious advantages, it is worth noting that above situations all have the same PS with 20B.

The proposed OGA method also has some advantages in terms of PS. We change the PS for end devices and measure the corresponding PDR. From the Fig. 9, it can be seen that the PDR difference between different methods is not obvious when the payload is smaller as the gateway is not saturated. As the PS increases and the gateway becomes saturated, however, increasing the PS decreases the capacity, as expected. Specially, for 200 devices, the PDRs achieved for payload size of 80 are improved about 70%, which is quite significant.

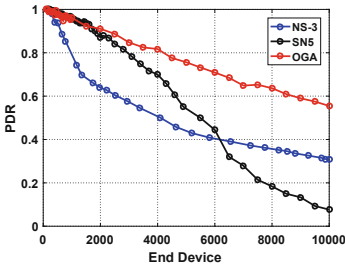


Fig. 8. Different dynamic parameter settings of PDR under 10000 device.

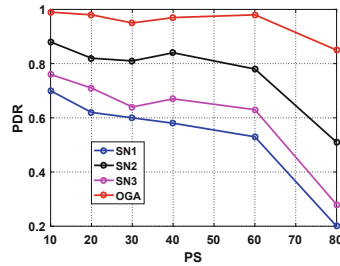


Fig. 9. Different PS of the PDR with 200 devices.

7 Conclusions

In this paper, The OGA method is introduced to solve the established channel collision and link budget model, which made the parameters of LoRa network selected dynamically. By the above method, the minimized collision probability parameter combination of which probability is 30% higher than the static parameter could be obtained. It is verified that LoRa network can be well scale by dynamic parameter selection. LoRa can support the deployment of a large number of applications, such as smart meters, smart parking and street lighting. However, most support multiple applications over a single current network deployment studies only support network simulation using a single IoT application, and different IoT applications have different data requirements on the amount of data to be transmitted and quality of service requirements. So our next step is to study application's data generation rate for different applications.

References

1. Da Xu, L., He, W., Li, S.: Internet of things in industries: a survey. *IEEE Trans. Ind. Inform.* **10**(4), 2233–2243 (2014)

2. Gubbi, J., Buyya, R., Marusic, S., et al.: Internet of Things (IoT): a vision, architectural elements and future directions. *Future Gen. Comput. Syst.* **29**(7), 1645–1660 (2013)
3. Augustin, A., Yi, J., Clausen, T., et al.: A study of LoRa: long range and low power networks for the internet of things. *Sensors* **16**(9), 1466–1475 (2016)
4. Wang, H., Fapojuwo, A.O.: A survey of enabling technologies of low power and long range machine-to-machine communications. *IEEE Commun. Surv. Tutorials* **19**(4), 2621–2639 (2017)
5. Tome, M., Nardelli, P., Alves, H., et al.: Long range low power wireless networks and sampling strategies in electricity metering. *IEEE Trans. Ind. Electron.* **66**(2), 1629–1637 (2019)
6. Wu, F., Wu, T., Yuce, M.: An internet-of-things (IoT) network system for connected safety and health monitoring applications. *Sensors* **19**(1), 1–21 (2019)
7. Kim, S., Yoo, Y.: Contention-aware adaptive data rate for throughput optimization in LoRaWAN. *Sensors* **18**(6), 1716–1732 (2018)
8. Cuomo, F., Campo, M., Caponi, A., et al.: EXPLoRa: extending the performance of LoRa by suitable spreading factor allocations. In: *Wireless and Mobile Computing, Networking and Communications (WiMob)*, March 2017, pp. 1–8 (2017)
9. Alonso, J.M., Alonso, J.M., Alonso, J.M., et al.: Do LoRa low-power wide-area networks scale? In: *ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems*, March 2016, pp. 59–67 (2016)
10. LoRa Alliance.LoRa. [OL], 25 June 2018. <https://www.lora-alliance.org/>
11. Peng, Y., Shangguan, L., Hu, Y., et al.: PLoRa: a passive long-range data network from ambient LoRa transmissions. In: *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, March 2018, pp. 147–160. ACM (2018)
12. Voigt, T., Bor, M., Roedig, U., et al.: Mitigating inter-network interference in lora networks. [arXiv:1611.00688](https://arxiv.org/abs/1611.00688) (2016)
13. Cuomo, F., Gamez, J.C.C., Maurizio, A., et al.: Towards traffic-oriented spreading factor allocations in LoRaWAN systems. In: *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, March 2018, pp. 1–8 (2018)
14. Bor, M., Roedig, U.: LoRa transmission parameter selection. In: *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, March 2017, pp. 27–34 (2017)
15. Adelantado, F., Vilajosana, X., Tuset-Peiro, P., et al.: Understanding the limits of LoRaWAN. *IEEE Commun. Mag.* **55**(9), 34–40 (2017)
16. Bouguera, T., Diouris, J.F., Chaillout, J., et al.: Energy consumption model for sensor nodes based on LoRa and LoRaWAN. *Sensors* **18**(7), 2104–2127 (2018)
17. Mikhaylov, K., Petajajarvi, J., Janhunen, J.: On LoRaWAN scalability: empirical evaluation of susceptibility to inter-network interference. In: *2017 European Conference on Networks and Communications (EuCNC)*, March 2017, pp. 1–6. IEEE (2017)
18. Petajajarvi, J., Mikhaylov, K., Pettissalo, M., et al.: Performance of a low-power wide-area network based on LoRa technology: doppler robustness, scalability, and coverage. *Int. J. Distrib. Sens. Netw.* **13**(3), 1–16 (2017)
19. Slabicki, M., Preamsankar, G., Di Francesco, M.: Adaptive configuration of LoRa networks for dense IoT deployments. In: *16th IEEE/IFIP Network Operations and Management Symposium (NOMS 2018)*, March 2018, pp. 1–9 (2018)
20. Van den Abeele, F., Haxhibeqiri, J., Moerman, I., et al.: Scalability analysis of large-scale LoRaWAN networks in ns-3. *IEEE Internet Things J.* **4**(6), 2186–2198 (2017)

21. Cattani, M., Boano, C., Romer, K.: An experimental evaluation of the reliability of lora long-range low-power wireless communication. *J. Sens. Actuator Netw.* **6**(2), 1–7 (2017)
22. Noreen, U., Bounceur, A., Clavier, L.: A study of LoRa low power and wide area network technology. In: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), March 2017, pp. 1–6 (2017)
23. Yong, W., Hui, L., Zixing, C.: An orthogonal design based constrained optimization evolutionary algorithm. *J. Eng. Optim.* **39**(6), 715–736 (2007)
24. Sanodiya, R.K., Saha, S., Mathew, J.: A kernel semi-supervised distance metric learning with relative distance: integration with a MOO approach. *Expert Syst. Appl.* **125**(7), 233–248 (2019)



A Weighted Voronoi Diagram Based Self-deployment Algorithm for Heterogeneous Mobile Sensor Network in Three-Dimensional Space

Li Tan^(✉), Xiaojiang Tang, Minghua Yang, and Haoyu Wang

Beijing Technology and Business University, Beijing 100048, China
tanli@th.btbu.edu.cn, tangxiaojiang_011@126.com,
yang_mh@qq.com, 1830401032@st.btbu.edu.cn

Abstract. For the node deployment problem in three-dimensional heterogeneous sensor networks, the traditional virtual force method is prone to local optimization and the parameters required for calculation are uncertain. A spatial deployment algorithm for 3D mobile wireless sensor networks based on weighted Voronoi diagram is proposed (TDWVADA) to solve the problem. Based on the positions and weights of all nodes in the monitoring area, a three-dimensional weighted Voronoi diagram is constructed. Next, the central position of each node's the Voronoi region is calculated and the position is regarded as target position of the node movement. Each node moves from the original position to the target position to complete one iteration. After multiple iterations, each node is moved to the optimal deployment location and network coverage is improved. In view of the initial centralized placement of sensor nodes, the addition of virtual force factors is added to the TDWVADA algorithm. An improved algorithm TDWVADA-I was proposed. The algorithm enables nodes that are centrally placed to spread quickly and speeds up deployment. The simulation results show that TDWVADA and TDWVADA-I effectively improve the network coverage of the monitored area compared to the virtual force algorithm and the unweighted Voronoi method. Compared with the virtual force method, the coverage of TDWVADA has increased from 90.53% to 96.70%, and the coverage of TDWVADA-I has increased from 81.12% to 96.56%. Compared with the Voronoi diagram method, the coverage of TDWVADA has increased from 85.01% to 96.70%, and the coverage of TDWVADA-I has increased from 80.82% to 96.56%. TDWVADA and TDWVADA-I also greatly reduce the energy consumption of the network. Experimental results demonstrate the effectiveness of the algorithms.

Keywords: Heterogeneous Wireless Sensor Networks (HWSNs) · 3D coverage · Area coverage · Voronoi diagram · Energy consumption

1 Introduction

The Internet of Things (IOT) is one of the most important research fields in the current information age. It has been widely used in all walks of life, such as smart home, smart city, smart community and vehicle network in the transportation field. As the bottom supporting part of the Internet of Things, wireless sensor networks (WSN) is one of the key technologies of the Internet of Things. Three-dimensional wireless sensor networks (3DWSNs) is a Wireless ad hoc network. It is composition by large number of micro-sensors deployed in the 3D monitoring area to extract energy from batteries. Three-dimensional wireless sensor networks have been successfully applied in many fields. Aguirre et al. applied it to real-time monitoring of urban traffic environment [1]. Jia et al. proposed an intelligent manhole cover management system based on edge computing by using vibration identification of sensors and narrow-band communication technology of Internet of things [2]. Manju et al. worked alternately by setting different nodes, and divided the priority of monitoring regions, increasing the priority of key monitoring areas in order to extend the life cycle of wireless sensor networks [3]. Fosalau et al. monitored natural disasters by deploying highly sensitive sensor nodes to perceive the direction and distance of soil movement [4].

Many research achievements have been made on the 3D deployment of wireless sensor network, Poduri et al. studied the feasibility of extending the existing 2D solution to 3D [5]. Huang et al. proposed a k coverage algorithm for monitoring target points, so that the target points are covered by at least k sensor nodes and improving the monitoring quality of the network [6]. The node activity scheduling solution proposed by Nauman et al. studied how to minimize the number of active nodes [7]. At the beginning, sensor nodes were randomly deployed to the object region, and the nodes that needed to be placed to the object location were allocated to the island activation state, while other nodes were switched to the dormant state to extend the entire network's face cycle.

The autonomous deployment of 3D wireless sensor networks has also achieved a lot. In literature [8], Li et al. proposed an autonomous deployment algorithm based on Voronoi partitioning principle for the k coverage problem of 3D surface. Brown et al. [9] studied the three-dimensional full space coverage of indoor wireless video sensor network by using heuristic greedy algorithm and enhanced depth-first algorithm, and obtained good experimental results. Temel, Akbarzadeh, Topcuoglu et al. [10–12] used heuristic algorithms to solve the problem of maximizing the coverage of three-dimensional terrain surfaces. Li et al. and Boufares et al. used virtual force strategy to autonomously deploy 3D mobile wireless sensor networks, and each sensor node could move in 3D target space according to the virtual force [13, 14]. Boufares et al. autonomously deployed the 3D plane using the virtual force strategy and completed the deployment of the 3D plane through the autonomous movement of nodes [15]. In literature [16], Yang et al. For the first time put forward the method of using discrete wavelet transform to detect the coverage cavities of 3D surface wireless sensor network and completed the node deployment of 3D surface by improving the artificial bee colony algorithm.

Above all, most of the existing deployment algorithms are based on random or deterministic strategies proposed for 3D static wireless sensor networks, some researchers have also proposed the deployment of 3D wireless sensor networks using virtual force method, but the virtual force may produce local optimality, leading to overlapping coverage and covering holes. This paper proposes a weighted Voronoi diagram method for autonomous deployment of 3D heterogeneous mobile sensor networks to solve the above problems.

2 Environmental Assumptions

This paper focuses on the problem of node deployment in a bounded 3D region, which can be abstracted as deployment in a cube region and regarded as a 3D target monitoring area. To ensure coverage and network connectivity in the monitoring area, the following assumptions are made.

A1: the monitoring area of wireless sensor network is an ideal area without obstacles and other adverse factors.

A2: the location information of each node in the wireless sensor network can be obtained by the positioning system.

A3: the sensor node perception model adopts the traditional binary perception model (0–1 perception model), that is, the probability of the node perceives the event within the perception range is 1, whereas the probability is 0.

A4: In this paper, the Cartesian coordinate system is adopted, and the node adopts the spherical model, with the node position as the center of the sphere. In the range of the node as the center of the sphere, the node can sense the surrounding environment, and the radius of the node is the center of the sphere. Within, nodes can communicate with each other, and the node model is shown in Fig. 1. The communication between nodes uses 5 g communication technology to realize long-distance communication and ensure the connectivity of the wireless sensor network. In this paper, all nodes are set to have the same parameters except for the perceived radius, which constitutes a three-dimensional heterogeneous sensor network.

3 Deployment Algorithm

In this paper, a novel space self-deployment algorithm for heterogeneous wireless sensor networks based on weighted Voronoi diagram is proposed, and an effective improvement is proposed for node centralization placement, so that the algorithm can adapt to different node placement situations. Both algorithms are described in detail in the following.

3.1 Voronoi Partitioning Algorithm

Voronoi diagram is a basic data structure about space partition. It uses Euclidean distance as a metric to divide the two-dimensional plane region or three-dimensional

space region. A Voronoi diagram of a three-dimensional point set is a spatial division of the point set. Each node after the division corresponds to a only polyhedral Voronoi region, which is the region closest to the node. The regions are seamlessly joined together without overlap to cover the three-dimensional space determined by the whole point set. As shown in Fig. 2, black is the node and red is the vertex of the Voronoi region polyhedron corresponding to the node.

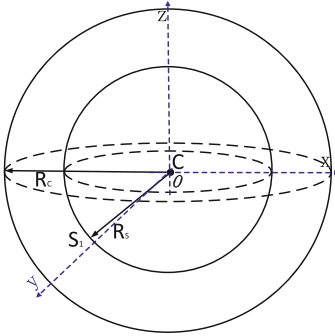


Fig. 1. Node model

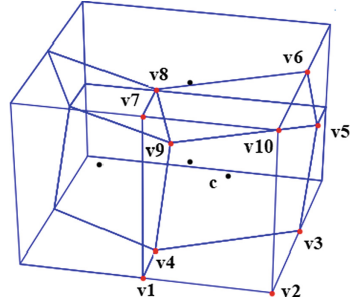


Fig. 2. Three-dimensional Voronoi diagram

3.2 Space Deployment Based on Weighted Voronoi Diagram Algorithm (TDWVADA)

According to the weights of each node, TDWVADA constructs a three-dimensional weighted Voronoi diagram in three-dimensional space to obtain the set of polyhedral vertices corresponding to each node, as shown in Fig. 2. The only polyhedron corresponding to node c is the enclosed polyhedron $\{v_1, v_2, v_3, v_4, v_5, v_6, v_7, v_8, v_9, v_{10}\}$. The center of the polyhedron is taken as the target position of the moving node, so that the position of the node can be constantly updated, and the node can be deployed in the whole three-dimensional space, and the coverage of the whole monitoring area can be improved. Suppose the location set of all nodes is $\mathbf{C}=\{C_1, C_2, \dots, C_n\}$, the set of perceived radius of nodes is $\mathbf{R}_s=\{R_{s_1}, R_{s_2}, \dots, R_{s_n}\}$, the set of polyhedral corresponding to each node in the weighted Voronoi diagram is $\mathbf{V}=\{V_1, V_2, \dots, V_n\}$, the center set of polyhedron corresponding to each node is $\mathbf{C}'=\{C'_1, C'_2, \dots, C'_n\}$, the weight set of nodes is $\mathbf{W}=\{w_1, w_2, \dots, w_n\}$, and the volume of the monitoring area is A . The weight of the current node C_i is w_1 , the perceived radius is R_{s_i} , the vertex set of the corresponding polyhedron is $V_i=\{v_1, v_2, \dots, v_k\}$, and the polyhedron center is C'_i . The weight of the node and the calculation formula of the polyhedron center are shown in formulas (1) and (2).

$$w_i = \sqrt[3]{\frac{3n^2A}{4\pi}} \times \frac{R_{s_i}}{\sum_{j=1}^n R_{s_j}} \tag{1}$$

$$(x_{C'_i}, y_{C'_i}, z_{C'_i}) = \left(\frac{\sum_{j=1}^n x_{v_j}}{n}, \frac{\sum_{j=1}^n y_{v_j}}{n}, \frac{\sum_{j=1}^n z_{v_j}}{n} \right) \quad (2)$$

Where $(x_{C'_i}, y_{C'_i}, z_{C'_i})$ is the coordinate of C'_i , $(x_{v_j}, y_{v_j}, z_{v_j})$ is the vertex coordinate of the polygon. The calculation formula of displacement vector of nodes \vec{M}_{v_i} is shown in formula (3):

$$\vec{M}_{v_i} = C'_i - C_i \quad (3)$$

The specific description of TDWVADA algorithm is as follows¹:

Input: Sensor nodes location C ; Monitoring area volume A ; Sensing radius of sensor nodes R_s ; The maximum number of iterations $Maxloop$

Output: Sensor nodes final location C'

```

/*      Initialization      */
1 Set  $Maxloop$ 
2 Randomly set sensor node initial position set  $C$ 
3 Set monitoring area volume  $A$ 
4 Set sensing radius of sensor nodes  $R_s$ 
5 Set  $n = 0$ 
6 Compute  $W$ 
/*      Main Loop      */
7 while  $n < maxloop$  do
8    $V \leftarrow WVoronoi(C, W, A, R_s)$ 
/* Construct a three-dimensional weighted Voronoi diagram */
9   for  $i=1$  to  $Length(V)$  do
10    Compute  $C'_i$ 
11     $\vec{M}_{v_i} \leftarrow C'_i - C_i$ 
12    Add  $C'_i$  to  $C'$ 
13    Add  $\vec{M}_{v_i}$  to  $M_v$ 
14   End for
15    $C \leftarrow C'$ 
16 End while

```

3.3 An Improved Algorithm for Nodes Set Placement (TDWVADA-I)

In the case of node centralization positioned, TDWVADA will cause local optimization and the deployment task of nodes cannot be completed well. An improved TDWVADA-I algorithm is proposed on the basis of TDWVADA so that centrally-placed nodes can

¹ In this paper, the generation algorithm of three-dimensional weighted Voronoi graph adopts the generation algorithm in the Voro++ library.

disperse rapidly. TDWVADA -I based on TDWVADA combines the principle of virtual force algorithm, that is between the various nodes and between nodes and border joined the virtual repulsion, virtual boundary will generate virtual repulsion to the nodes near the border, prevent nodes move beyond the monitoring area, and the same time each node will also be produced by the neighbor node repulsive action, whether there is virtual repulsion between nodes is determined according to the distance between nodes. The boundary of the monitored area generates virtual repulsive force or no virtual force according to the distance from the current node C_i . The nodes move under the action of virtual resultant force and Voronoi diagram so that the concentrated nodes can disperse.

Define the set of neighbor nodes of node C_i as $C_n = \{C_j | D(C_i, C_j) < D_{th}\}$, where $D(C_i, C_j)$ is the Euclidean distance calculation formula of node C_i and node C_j , as shown in formula (4), D_{th} is the range of repulsive forces generated between nodes, and the value of D_{th} in this paper is $(R_{s_i} + R_{s_j})$.

$$D(C_i, C_j) = \sqrt{(x_{C_j} - x_{C_i})^2 + (y_{C_j} - y_{C_i})^2 + (z_{C_j} - z_{C_i})^2} \quad (4)$$

The calculation formula of virtual repulsive force generated by neighbor nodes of node C_i on node C_i is shown in formula (5):

$$\vec{F}_{(i,j)} = \begin{cases} \left(\frac{k}{D_{(C_i, C_j)}^\lambda}, -\vec{\alpha}_{(i,j)} \right), & 0 < D(C_i, C_j) < D_{th} \\ \infty, & D(C_i, C_j) = 0 \\ \vec{0}, & other \end{cases} \quad (5)$$

Where k , λ is the repulsive force coefficient, $\vec{\alpha}_{(i,j)}$ is the unit vector, represents the direction from the node C_i to the neighbor node $C_j \in C_n$, D_{th} is the critical point generated by the virtual repulsive force, and the value D_{th} in this paper is $(R_{s_i} + R_{s_j})$. The virtual resultant force of neighbor node C_i is shown in formula (6) :

$$\vec{F}_{i_r} = \sum_{C_j \in C_n} \vec{F}_{(i,j)} \quad (6)$$

Where \vec{F}_{i_r} is the resultant of the repulsive force on node C_i , C_j is the neighbor node of node C_i , $\vec{F}_{(i,j)}$ is the virtual repulsive force on node C_i , and C_n is the set of neighbor nodes of node C_i . Node C_i is not only affected by the virtual repulsion of neighbor nodes, but also by the virtual repulsion from the boundary of the monitoring area, the virtual repulsion of the boundary to the node ensures that the node will not be deployed outside the monitoring area. The calculation of virtual repulsive force of each boundary on the node C_i is shown in formula (7)–(9). In formula (7)–(9), L_x and H_x are the boundary plane in the direction of X axis, L_y and H_y are the boundary plane in the direction of Y axis, L_z and H_z are the boundary plane in the direction of Z axis, and d_{th} are the critical distance to generate virtual repulsive force in the boundary plane. In this paper, the value is R_{s_i} , $\vec{\alpha}$ is the unit vector, and represents the direction from the node to the boundary plane.

$$F_{ix} = \begin{cases} \left(\frac{k}{(x_{C_i} - L_x)^2}, -\vec{\alpha}_{(i,L_x)} \right), & 0 \leq x_{C_i} - L_x < d_{th} \\ \left(\frac{k}{(H_x - x_{C_i})^2}, -\vec{\alpha}_{(i,H_x)} \right), & 0 \leq H_x - x_{C_i} < d_{th} \\ \vec{0}, & other \end{cases} \quad (7)$$

$$F_{iy} = \begin{cases} \left(\frac{k}{(y_{C_i} - L_y)^2}, -\vec{\alpha}_{(i,L_y)} \right), & 0 \leq y_{C_i} - L_y < d_{th} \\ \left(\frac{k}{(H_y - y_{C_i})^2}, -\vec{\alpha}_{(i,H_y)} \right), & 0 \leq H_y - y_{C_i} < d_{th} \\ \vec{0}, & other \end{cases} \quad (8)$$

$$F_{iz} = \begin{cases} \left(\frac{k}{(z_{C_i} - L_z)^2}, -\vec{\alpha}_{(i,L_z)} \right), & 0 \leq z_{C_i} - L_z < d_{th} \\ \left(\frac{k}{(H_z - z_{C_i})^2}, -\vec{\alpha}_{(i,H_z)} \right), & 0 \leq H_z - z_{C_i} < d_{th} \\ \vec{0}, & other \end{cases} \quad (9)$$

The resultant of virtual repulsive force \vec{F}_{ixyz} of all boundaries on node C_i is calculated by formula (10):

$$\vec{F}_{ixyz} = \vec{F}_{ix} + \vec{F}_{iy} + \vec{F}_{iz} \quad (10)$$

The resultant of all repulsive forces on the node C_i is the total resultant of the resultant force on the node and the resultant force on the node on the boundary. The calculation formula is shown in formula (11), and the calculation of the movement vector of the node under the action of virtual force is shown in formula (12).

$$\vec{F}_i = \vec{F}_{i_r} + \vec{F}_{ixyz} \quad (11)$$

$$\vec{M}_{F_i} = \lambda \vec{F}_i \quad (12)$$

λ is the moving coefficient in formula (12). Then, the resultant displacement vector of the current node C_i is calculated according to the \vec{M}_{V_i} sum in formula (2) and \vec{M}_{F_i} in formula (12). The calculation method is shown in formula (13). The resultant displacement vector determines the moving distance and direction of the node.

$$\vec{M}_i = \mu \vec{M}_{F_i} + \vec{M}_{V_i} \quad (13)$$

μ in the formula (13) is to adjust the coefficient of the influence of the virtual force moving vector on the whole moving vector, due to the method of weighted Voronoi diagram is not ideal for the deployment of centrally placed nodes, but the deployment

effect after the nodes are dispersed is better than the virtual force method, so by constantly adjusting the movement vector of the virtual force changes the impact on the moving vector of virtual force, this enables the node deployment to achieve an ideal effect, in this article, calculation formula of μ is as shown in formula (14):

$$\mu = \delta^{\varepsilon*i} \quad (14)$$

In Eq. (14), the value of δ is less than 1, and the value of ε is greater than 1, The two parameters determine the value μ together.

The specific description of TDWVADA-I algorithm is as follows:

Algorithm2: TDWVADA-I

Input: Sensor nodes location C ; Monitoring area volume A ; Sensor nodes perceive radius R_s ; The maximum number of iterations $Maxloop$

Output: Sensor nodes final location C'

```

/*      Initialization      */
1 Set  $Maxloop$ 
2 Centrally and randomly set the sensor node initial position set  $C$ 
3 Set monitoring area  $A$ 
4 Set sensor nodes perceive radius  $R_s$ 
5 Set  $n = 0$ 
6 Compute  $W$ 
/*      Main Loop      */
7 while  $n < maxloop$  do
8    $V \leftarrow WVoronoi(C, W, A, R_s)$ 
   /*Construct a three-dimensional weighted Voronoi diagram*/
9   for  $i = 1$  to Length( $C$ ) do
10    Compute  $C_i'$ 
11     $\vec{M}_{v_i} \leftarrow C_i' - C_i$ 
12    Add  $\vec{M}_{v_i}$  to  $M_v$ 
13   End for
14   for  $i = 1$  to Length( $C$ ) do
15    Compute  $\vec{F}_{i_{XYZ}}$ 
16    Compute  $\vec{F}_{i_r}$ 
16     $\vec{F}_i \leftarrow \vec{F}_{i_{XYZ}} + \vec{F}_{i_r}$ 
17     $\vec{M}_{F_i} \leftarrow \lambda \vec{F}_i$ 
18    Add  $\vec{M}_{F_i}$  to  $M_F$ 
19   End for
20    $M \leftarrow M_v + \mu M_F$ 
21    $C' \leftarrow C + M$ 
22    $C \leftarrow C'$ 
23 End while

```

The Voronoi diagram of 3D point set has the worst algorithm complexity of $O(n^2)$. Assuming that the number of iterations of the algorithm in this paper is m , the time complexity of the algorithm TDWVADA is $O(mn^2)$. The time complexity of virtual force algorithm the maximum is $O(n^2)$, generally less than $O(n^2)$, so the time complexity of TDWVADA-I algorithm the least is $O((m + 1)n^2)$.

4 Simulation and Analysis

4.1 Simulation

This experiment adopts Python2.7 simulation experiment platform, monitoring area set to 50 m * 50 m * 50 m of the three-dimensional space, the node's perception radius R_s is a random integer between 3 to 8 m, the number of nodes deployed in monitoring area is 300, the initial node of algorithm TDWVADA random distribution in the whole monitoring area, the initial node of algorithm TDWVADA -I random distribution in the center of the monitoring area of 10 m *10 m * 10 m, that is the values of x, y, z is in the areas of 20 to 30. The specific parameters are shown in Table 1.

Table 1. Parameter values table

Parameters	Values
Area size A	50 m * 50 m * 50 m
Nodal number n	300
Random deployment area	[(0,50), (0,50), (0,50)]
Centralized deployment area	[(20,30), (20,30), (20,30)]
δ	0.9
ε	2

The deployment of algorithm TDWVADA and algorithm TDWVADA-I are shown in Figs. 3 and 4, respectively showing the distribution of nodes in the initial state, three iterations and 50 iterations.

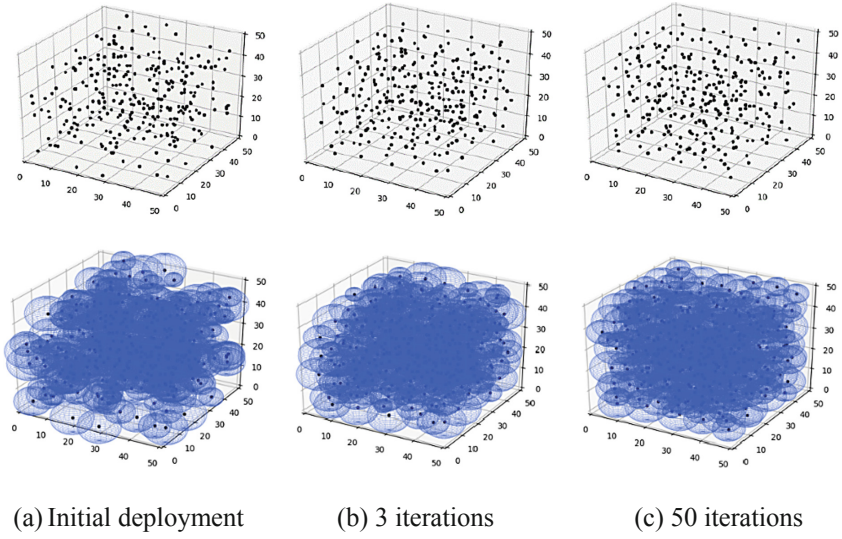


Fig. 3. TDWVADA deployment

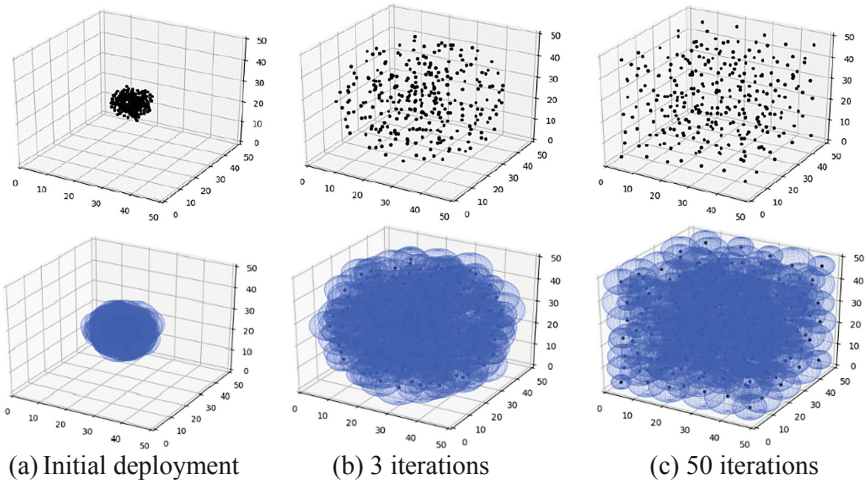


Fig. 4. TDWVADA-I deployment

4.2 Results Analysis

In wireless sensor networks, the main indicators to evaluate the performance of the algorithm are as follows: (1) coverage rate: one of the important indicators to evaluate the performance of the sensor network, which reflects the comprehensive monitoring ability of the sensor network to the monitored objects; (2) rate of convergence: refers to the time required for the coverage rate to reach a relatively stable state in the whole

deployment process. The smaller the time is, the faster the rate of convergence will be and the better the performance will be. (3) energy consumption: energy consumption can directly affect the service life of wireless sensor network. In this paper, the average moving distance of nodes is used to reflect energy consumption. If the average moving distance of wireless sensor network is lower during deployment, the energy consumption of the network will be lower.

Table 2. Algorithm comparison

Method of deployment	Algorithm	Coverage	Time cost (s)	Average moving distance (m)
Initial random deployment	VFA	90.53%	22.298	156.29
	Voronoi	85.01%	3.285	5.717
	Voronoi+VFA	85.33%	6.181	18.27
	TDWVADA	96.70%	3.230	13.08
	TDWVADA-I	94.81%	6.303	22.09
Initial centralized deployment	VFA	81.12%	21.641	234.81
	Voronoi	76.28%	3.325	16.34
	Voronoi+VFA	80.82%	6.172	29.38
	TDWVADA	84.12%	5.028	23.71
	TDWVADA-I	96.65%	6.311	37.61

Table 2 compares the coverage rate, the time spent and the average movement distance of each node when the algorithm (VFA) based on virtual force, the algorithm (Voronoi) based on Voronoi diagram and the algorithm combined with virtual force (Voronoi+VFA), TDWVADA and TDWVADA-I first iterated for 50 times under different initial conditions. It can be seen from Table 2 that the coverage rate of TDWVADA and TDWVADA-I in the case of initial random deployment nodes is much higher than that of other algorithms. The time consumed by VFA is 7 times bigger than that of TDWVADA, 3.5 times bigger than that of TDWVADA-I, and the average moving distance is 12 times bigger than that of TDWVADA and 7 times bigger than that of TDWVADA-I. Compared with TDWVADA-I, TDWVADA has more advantages in time spent and moving distance. In the case of initial centralized deployment, the coverage rate of TDWVADA-I reaches 96.65%, which is much higher than other algorithms. Compared with VFA algorithm, in terms of the time consumed by the algorithm and the moving distance, it also has a great improvement. For the unweighted Voronoi diagram method, the Voronoi region for all the nodes what will eventually occur is not very different in size, but this area cannot be completely covered for the nodes with a small perceptual range, it will cause cover hole. To the nodes with a larger perceptual range, the node will have a large coverage beyond the Voronoi region of the node, overlapping, and can't do deployment task well. For the VFA method, due to the interaction force between nodes, all nodes will be concentrated in one area, resulting in the overlapping of coverage in this area. At the same time, for some areas close to the boundary, the coverage hole will appear and the deployment task cannot be well completed.

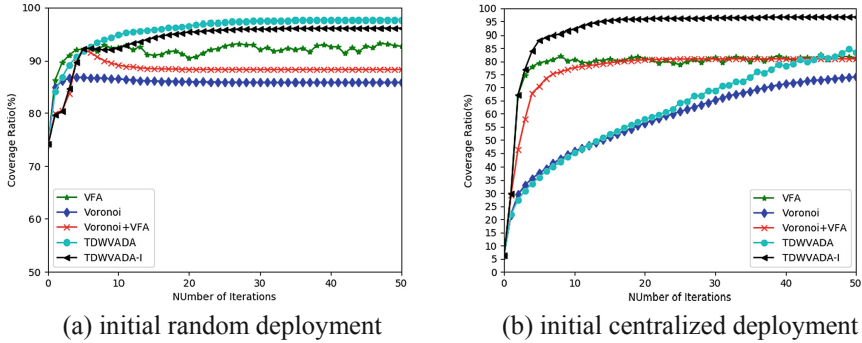


Fig. 5. The relationship between the number of iteration and the coverage ratio

In order to study the convergence speed of the algorithm, 300 nodes were deployed in the target area, and the relationship between the number of iterations and the coverage rate was simulated. The results are shown in Fig. 5. From Fig. 5 shows that in the case of random deployment points, TDWVADA and TDWVADA-I after 10 times iteration the coverage have exceeded other algorithms, after 30 times loop coverage will basic stable, and reached more than 95%, Voronoi and Voronoi+VFA convergence is faster but the coverage is lower than under TDWVADA and TDWVADA-I, VFA coverage reached 90% after 5 times iterations, but after this the coverage can not reach a steady state, always fluctuates up and down at 90%. This shows that it is difficult for the VFA algorithm to make the whole sensor network reach a relatively stable state. Under the condition of the centralized deployment, TDWVADA-I after 20 times iteration, the coverage reached over 95%, although TDWVADA can also achieve a higher level of coverage finally, but the convergence speed is too slow, from Fig. 5(b) can find the Voronoi and TDWVADA after 50 times iteration are still no convergence, VFA and the Voronoi algorithm convergence needed, though less iteration times but coverage far less than TDWVADA algorithm. According to the running time of the algorithm, the convergence time of TDWVADA and TDWVADA-I is still lower than that of VFA algorithm.

Node energy consumption is very important for wireless sensor networks, the average moving distance can reflect the energy saving effect of the algorithm to a certain degree, Fig. 6 shows the 300 nodes 50 times each iteration each node in the process of point moving average distance, the ordinate adopted index coordinates in the graph, Fig. 6(a) is random deployment, Fig. 6(b) shows the centralized deployment.

Can be found from the Fig. 6, both random deployment and centralized deployment, eventually the average moving distance of the nodes in VFA algorithm is always from 3 m to 4 m, at the same time can be found from the Fig. 5, coverage of the VFA algorithm is also fluctuates around 90%, this shows again that VFA algorithm is difficult to make the whole network to achieve a relatively stable state, the average distance of other algorithms is quickly dropped to below 1 m, when the network achieve a relatively stable, the average movement distance gradually tend to be zero. However, in Fig. 6(b), the average moving distance of the algorithm TDWVADA does not approach to 0 in the end because the algorithm has not yet converged and the

wireless sensor network has not in a stable state. According to the experimental results of average moving distance, the energy saving effects of TDWVADA and TDWVADA-I is very obvious.

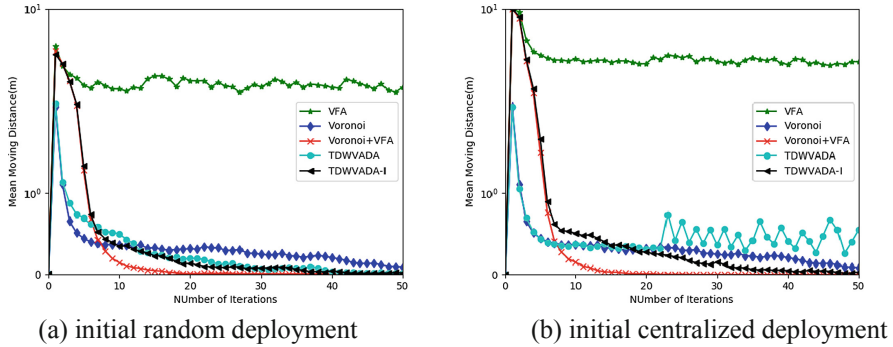


Fig. 6. The relationship between the number of iteration and average moving distance

In Fig. 7, the total movement distance, total consumption time and final coverage rate of the five algorithms after 50 times iterations were compared under the initial random deployment and centralized deployment, Comprehensive the three factors of energy saving, time efficiency and coverage, at the initial random deployment, the algorithm TDWVADA has the best performance, during initial centralized deployment, algorithm TDWVADA-I has the best performance. Figure 8 studies the influence of different deployment locations on the TDWVADA-I algorithm during the initial nodal centralized deployment. In the experiment, four types of node deployment locations are selected. The First type is close to three boundary planes, in the experiment, the values of x , y and z are selected from the $10\text{ m} * 10\text{ m} * 10\text{ m}$ area between 0 and 10 m, as shown in the First area in Fig. 9. The Second type is the area close to the two boundary planes. The values of x and y are selected in the experiment between 0 and 10 m, and z is between 20 and 30 m in the area of $10\text{ m} * 10\text{ m} * 10\text{ m}$, as shown in the Second area in Fig. 9. The third category is the area close to a boundary plane. In the experiment, the values of x are all between 0 and 10 m, and the values of y and z are within the $10\text{ m} * 10\text{ m} * 10\text{ m}$ between 20 and 30 m, as shown in the third region in Fig. 9. The fourth category is the region not close to the boundary plane. In the experiment, the value of x , y and z are selected from the $10\text{ m} * 10\text{ m} * 10\text{ m}$ region between 20 m and 30 m, as shown in the fourth region in Fig. 9. Can be seen from the Fig. 8, the location of the initial node set deployment for algorithm TDWVADA-I have influence, the convergence speed and average moving distance of the algorithm are optimal when the initial node is deployed in the fourth region, these are the Third, Second, and the First area in the region. Therefore, in the actual deployment, it is necessary to deploy as far as possible in the center of the monitoring area to improve the performance of the algorithm.

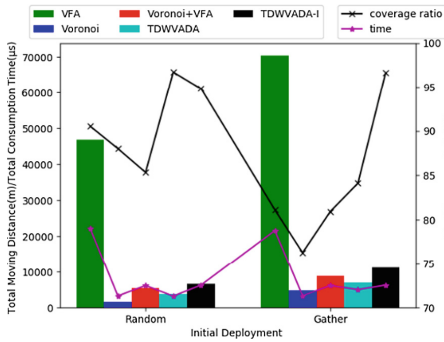


Fig. 7. Algorithm comparison

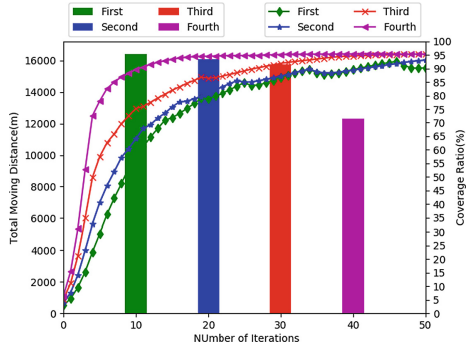


Fig. 8. The influence of position on Algorithm

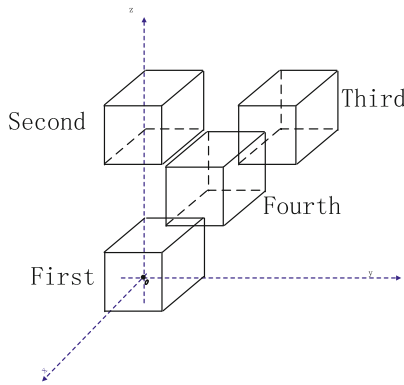


Fig. 9. Initial node centralized deployment area

Can be seen from the above analysis of the proposed TDWVADA algorithm and TDWVADA-I, the performance of the algorithm is superior to other algorithms, TDWVADA-I algorithm can achieve higher coverage in both random deployment of initial node and centralized deployment of initial node, but in the case of the initial nodes random deployment the convergence speed and energy consumption are not as good as TDWVADA algorithm; the performance of TDWVADA-I algorithm is much better than that of TDWVADA algorithm in the case of initial node-set deployment.

5 Conclusions

In this paper the deployment of three dimensional heterogeneous wireless sensor networks are studied, an algorithm for spatial autonomous deployment of heterogeneous mobile sensor networks based on three-dimensional weighted Voronoi diagram is proposed, and on this basis an improved algorithm TDWVADA-I is proposed for the centralized placement of sensor nodes, and how to achieve the highest coverage

through autonomous deployment of nodes in 3D monitoring area is discussed in depth, through the simulation analysis of the rate of convergence of the algorithm and the energy consumption of the sensor network problems. Experimental results show that the proposed algorithm can greatly improve the coverage of heterogeneous wireless sensor networks and has a very good performance in energy saving.

Acknowledgement. This research was funded by the National Natural Science Foundation of China grant number (61702020), Beijing Natural Science Foundation grant number (4172013) and Beijing Natural Science Foundation-Haidian Primitive Innovation Joint Fund grant number (L182007).

References

1. Aguirre, E., Lopez-Iturri, P., Azpilicueta, L., et al.: Design and implementation of context aware applications with wireless sensor network support in urban train transportation environments. *IEEE Sens. J.* **17**(1), 169–178 (2017)
2. Jia, G., Han, G., Rao, H., et al.: Edge computing-based intelligent manhole cover management system for smart cities. *IEEE Internet Things J.* **PP**(99), 1 (2017)
3. Manju, Chand, S., Kumar, B.: Maximising network lifetime for target coverage problem in wireless sensor networks. *IET Wirel. Sens. Syst.* **6**(6), 192–197 (2016)
4. Fosalau, C., Zet, C., Petrisor, D.: Implementation of a landslide monitoring system as a wireless sensor network. In: *Ubiquitous Computing, Electronics & Mobile Communication Conference*, pp. 1–6. IEEE (2016)
5. Radani, Z.M., Samavi, S., Fooladgar, F.: Multi plane volumetric coverage in wireless visual sensor network. In: *Electrical Engineering*, pp. 758–762. IEEE (2012)
6. Huang, C.-F., Tseng, Y.-C., Lo, L.-C.: The coverage problem in three-dimensional wireless sensor networks. *J. Interconnection Netw.* **8**(3), 209–227 (2007)
7. Nauman, A.: Optimizing coverage in 3D wireless sensor networks. In: *Smart Wireless Sensor Networks*, pp. 189–204 (2010)
8. Li, F., Luo, J., Wang, W., et al.: Autonomous deployment for load balancing, -surface coverage in sensor networks. *IEEE Trans. Wirel. Commun.* **14**(1), 279–293 (2015)
9. Brown, T., Wang, Z., Shan, T., et al.: On wireless video sensor network deployment for 3D indoor space coverage. In: *Southeastcon*, pp. 1–8. IEEE (2016)
10. Temel, S., Unaldi, N., Kaynak, O.: On deployment of wireless sensors on 3-D terrains to maximize sensing coverage by utilizing cat swarm optimization with wavelet transform. *IEEE Trans. Syst. Man Cybern. Syst.* **44**(1), 111–120 (2013)
11. Akbarzadeh, V., Gagne, C., Parizeau, M., et al.: Probabilistic sensing model for sensor placement optimization based on line-of-sight coverage. *IEEE Trans. Instrum. Measur.* **62**(2), 293–303 (2013)
12. Topcuoglu, H.R., Ermis, M., Sifyan, M.: Positioning and utilizing sensors on a 3-D terrain part ii—solving with a hybrid evolutionary algorithm. *IEEE Trans. Syst. Man Cybern. Part C* **41**(4), 470–480 (2011)
13. Li, X., Ci, L., Yang, M., Tian, C., Li, X.: Deploying three-dimensional mobile sensor networks based on virtual forces algorithm. In: Wang, R., Xiao, F. (eds.) *CWSN 2012. CCIS*, vol. 334, pp. 204–216. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-36252-1_19
14. Boufares, N., Khoufi, I., Minet, P., et al.: Three dimensional mobile wireless sensor networks redeployment based on virtual forces. In: *2015 International Wireless*

- Communications and Mobile Computing Conference (IWCMC), Dubrovnik, pp. 563–568. IEEE (2015)
15. Boufares, N., Minet, P., Khoufi, I., et al.: Covering a 3D flat surface with autonomous and mobile wireless sensor nodes. In: 2017 13th International Wireless Communications and Mobile Computing Conference (IWCMC), Valencia, pp. 1628–1633. IEEE (2017)
 16. Yang, H., Li, X., Wang, Z., et al.: A novel sensor deployment method based on image processing and wavelet transform to optimize the surface coverage in WSNs. *Chin. J. Electron.* **25**(3), 495–502 (2016)



Joint Uplink and Downlink Optimization for Resource Allocation Under D2D Communication Networks

Di He, Guangsheng Feng^(✉), Bingyang Li, Hongwu Lv, Huiqiang Wang,
and Quanming Li

College of Computer Science and Technology, Harbin Engineering University,
Harbin 150001, China
fengguangsheng@hrbeu.edu.cn

Abstract. We study the joint uplink and downlink (JUAD) resource allocation problem in D2D networks, where D2D sender communicates with D2D recipient by reusing the channel of cellular users (CUs). In order to maximize the throughput of D2D networks, we model the JUAD problem as a mixed integer nonlinear programming problem (MINLP). Since the problem is NP hard, to solve it better, we divide it into two sub-problems by analyzing the structure of the primal problem, including channel assignment and power allocation. Then, we turn the sub-problem of power allocation into convex problem by the Lagrangian dual theory for getting the optimal power value of CUs and D2D pair. Next, an improved Hopcroft-Karp algorithm is proposed to solve the sub-problem of channel allocation, which has lower complexity compared with the traditional channel allocation approaches. Finally, extensive simulations show that our proposed approach achieves a near optimal solution.

Keywords: Device to Device communication · Power allocation · Joint uplink and downlink resource allocation · Hopcroft-Karp algorithm

1 Introduction

The exponential growth of smart devices and corresponding applications leads to an increasing demand on high data rate access, which can hardly be accommodated under traditional wireless communication techniques [1]. To efficiently utilize the spectrum resource as well as to alleviate the huge infrastructure investment of operators, D2D communications allow two nearby end users to communicate directly instead of communicating via a base station, which provide the following potential gains: (i) the reuse gain due to the sharing spectrum resource with other cellular users (CUs) [2], (ii) the proximity gain due to the good channel condition of D2D pairs in proximity [3], and (iii) the hop gain due to one-hop communication between two user devices (UEs) [4].

In D2D communications underlying cellular networks, there are two ways for D2D users to access the spectrum: overlay spectrum sharing and underlying

spectrum sharing [5]. In the former one, a base station (BS) allocates a portion of the idle spectrum for D2D communications, obviously, which cannot effectively solve the shortage problem of BS spectrum resources [6]. Besides, the quality of service (QoS) for CUs may be reduced due to the increasing number of D2D pairs in wireless networks. In the latter one, each D2D pair can share subcarrier resources with CUs, which can greatly improve the spectrum efficiency and ensure the QoS of cellular communications [7]. Thus, in this paper, we focus on the underlying spectrum sharing problem where each D2D pair can reuse the spectrum resources of CUs.

However, due to the coexistence of D2D pairs and CUs that operate under the same spectrum, the D2D pair will cause some interference to CUs and BS [8]. In addition, the CU may also interfere with the D2D pair, which greatly reduce the communication rate of D2D pair. Since the source of interference depends on the shared resource in uplink (UL) or downlink (DL) phases [9], it is crucial to devise efficient resource allocation management to maximize the overall throughput of D2D networks.

In recent years, many resource allocation methods in UL and DL phases are proposed in literatures [10–14]. In [10], the authors studied the joint optimization of subcarrier allocation, uplink and downlink user pairs and power allocation to maximize the overall throughput. In [11], the authors first adopted Kuhn-Munkras (KM) algorithm to solve the problem of channel allocation, which the complexity of KM is increasing with the number of users. By contrast, in [12], the authors proposed an improved Hungarian algorithm to reduce the complexity of channel allocation. But it only assign a subcarrier to each D2D pair. Then a novel scheme that allow each D2D pair to reuse channels of multiple CUs (D2D-CUs) was proposed in [13]. However, in [10–13], the authors did not consider the impact of interference. In [14], the authors first proposed a packet delivery mechanism to reduce the required bandwidth in UL and DL scheme, but its objective is not the problem of maximizing throughput of D2D networks.

Different from the above works, we design a joint uplink and downlink (JUAD) resource allocation scheme in frequency division duplex (FDD) system, where each D2D pair can reuse the channels of multiple CUs. The main contributions of this paper are summarized as follows:

- i. We formulate the JUAD problem as a MINLP problem with the aim of maximizing the overall throughput of D2D networks, in which each D2D pair is allowed to reuse the resources of multiple CUs.
- ii. We develop a two-step method to decouple the JUAD problem into two sub-problems, including power allocation and channel allocation. Then, we adopt a convex optimization technology and propose an improved Hopcroft-Karp algorithm to solve the power allocation problem and the channel allocation problem, respectively.
- iii. The simulation results highlight the effectiveness of our approach. By carefully checking all possible assignments between D2D pairs and CUs, the overall throughput of D2D networks can be improved.

2 System Model and Problem Statement

We consider a D2D-enabled system in a fully loaded single cell. As shown in Fig. 1, there are \mathcal{M} CUs coexisting with \mathcal{N} D2D pairs, denoted by set $\mathcal{M} = \{1, 2, \dots, m\}$ and set $\mathcal{N} = \{1, 2, \dots, n\}$, in which the number of CUs is greater than that of D2D pairs in cellular networks. The system employs the universal frequency reuse scheme, where BS pre-assigns an orthogonal uplink channel and an orthogonal downlink channel to each CU for ensuring the QoS of CUs [13]. Besides, giving the shortage of spectrum resource, each CU shares channel with only one D2D pair while each D2D pair can reuse channels of multiple CUs to transmit data. In order to guarantee that all D2D pairs can share channels with multiple CUs, we limit the number of reused channels of the D2D pair j by the quota q_j .

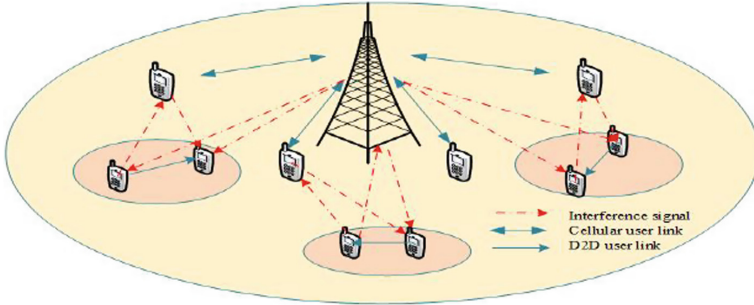


Fig. 1. System scenario of the device-to-device underlying communication.

We denote by $\mathbf{a}^u = \{a_{ij}^u\}$ and $\mathbf{a}^d = \{a_{ij}^d\}$ the uplink and downlink assignment vector of D2D pair, respectively, i.e., $a_{ij}^u = 1$ when D2D pair j reuses the uplink channel of the CU i , and otherwise $a_{ij}^u = 0$. Similarly, $a_{ij}^d = 1$ when D2D pair j reuses the downlink channel of the CU i , and otherwise $a_{ij}^d = 0$.

Since each CU shares channel with only one D2D pair and each D2D pair can share channel with a maximum number of CUs, channel assignment should satisfy

$$\sum_{i \in \mathcal{M}} (a_{ij}^u + a_{ij}^d) \leq q_j, \forall j \in \mathcal{N}, \quad (1)$$

$$0 \leq \sum_{j \in \mathcal{N}} a_{ij}^u \leq 1, \forall i \in \mathcal{M}, \quad (2)$$

$$0 \leq \sum_{j \in \mathcal{N}} a_{ij}^d \leq 1, \forall i \in \mathcal{M}. \quad (3)$$

Given that each D2D pair cannot choose to reuse both the uplink and the downlink resources of one CU, channel assignment also should satisfy

$$\left(\sum_{i=1}^m a_{ij}^u\right) \cdot \left(\sum_{i=1}^m a_{ij}^d\right) = 0, \forall j \in \mathcal{N}. \quad (4)$$

2.1 Transmission Rate

In uplink (UL) and downlink (DL) phases, since the impact of interference on CUs and D2D pair is different, we should derive the transmission rate formulas of CU and D2D pair in both UL and DL phases, respectively. Then the transmission rates of CUs and DUs can be calculated.

In UL communication phase, the communication of CU will cause interference to D2D pair. So the transmission rate of CU i , D2D pair j , and the rate constraint can be expressed as

$$R_i^{ul,c} \geq \gamma_{th}^c, \quad (5)$$

$$R_{ij}^{ul,d} \geq \gamma_{th}^d, \quad (6)$$

$$R_i^{ul,c} = \text{Blog}(1 + \psi_i^{ul,c}), \quad (7)$$

$$R_{ij}^{ul,d} = \text{Blog}(1 + \psi_{i,j}^{ul,d}), \quad (8)$$

where $\psi_i^{ul,c}$ and $\psi_{i,j}^{ul,d}$ are respectively the signal to interference plus noise ratio (SINR) of CU i and the SINR of D2D pair j , γ_{th}^c and γ_{th}^d are denoted by the uplink rate requirements of CU i and D2D pair j , the specific SINR formula of CU i and D2D pair j is shown below

$$\psi_i^{ul,c} = \frac{P_i^c \cdot h_{iB}}{\sum_{j=1}^n a_{ij}^u \cdot P_{ij}^d \cdot h_{jB} + N_0}, \quad (9)$$

$$\psi_{i,j}^{ul,d} = \frac{P_{ij}^d \cdot h_{ij}}{\sum_{j=1}^n a_{ij}^u \cdot P_i^c \cdot h_{ij} + N_0}, \quad (10)$$

where N_0 represents additive gaussian noise, h_{iB} and h_{ij} are respectively the channel gain between CU i and BS, the channel gain between CU i and D2D pair j , the transmission power of CU i and the transmission power of D2D pair j on uplink channel i are respectively denoted by P_i^c , P_{ij}^d . Due to the limited power of devices, the power value of devices should meet the following constraints

$$0 \leq P_i^c \leq P_i^{max}, \quad (11)$$

$$0 \leq \sum_{i \in M} \sum_{j \in N} (a_{ij}^u \cdot P_{ij}^d) \leq P_{total}^d, \quad (12)$$

where P_i^{max} represents the maximum power value of CU i , P_{total}^d represents the maximum power value of D2D pair.

In DL communication phase, BS will interfere with D2D pair, thus the SINR of CU i and D2D pair j are expressed as

$$\psi_i^{dl,c} = \frac{P_{Bi} \cdot h_i}{\sum_{j=1}^n a_{ij}^d \cdot P_{ij}^d \cdot h_{Bj} + N_0}, \quad (13)$$

$$\psi_{ij}^{dl,d} = \frac{P_{ij}^d \cdot h_{jj}}{\sum_{j=1}^n a_{ij}^d \cdot P_{Bi} \cdot h_{Bj} + N_0}, \quad (14)$$

where h_{jj} represents the channel gain between the D2D transmitter and receiver in D2D pair j , P_{Bi} represents the transmission power value of BS.

According to the formula of shannon, the transmission rate of CU i and D2D pair j in the downlink communication phase can be expressed as $R_i^{dl,c} = \text{Blog}(1 + \psi_i^{dl,c})$, $R_{ij}^{dl,d} = \text{Blog}(1 + \psi_{ij}^{dl,d})$. Similar to the uplink communication phase, the communication rate of DL phase also needs to meet the following constraints

$$R_i^{dl,c} \geq \gamma_{th}^c, \quad (15)$$

$$R_{ij}^{dl,d} \geq \gamma_{th}^d. \quad (16)$$

2.2 Problem Formulation

In this paper, we investigate a JUAD resource allocation problem to maximize the overall throughput of D2D networks while ensuring the QoS of both CUs and D2D pairs. Specifically, the optimization problem is as follows:

$$P1 : \max_{\mathbf{a}^u, \mathbf{a}^d, \mathbf{P}^d, \mathbf{P}^c} \sum_{i=1}^M \sum_{j=1}^N (a_{ij}^u \cdot R_{ij}^{ul,d} + a_{ij}^d \cdot R_{ij}^{dl,d}) \quad (17)$$

$$s.t. (1) - (6), (11), (12), (15), (16).$$

The problem P1 includes both continuous and discrete variables, which is a mixed integer nonlinear programming (MINLP) and usually mathematically intractable. To solve P1, we propose a JUAD resource allocation algorithm that includes power allocation and channel allocation.

3 Problem Solution

The JUAD resource allocation algorithm decomposes P1 into two sub-problems, including the problems of power allocation and channel allocation. First, we convert the power allocation problem to a convex problem by the Lagrangian dual theory and use the gradient descent algorithm to get the optimal powers of D2D pair and CU. Then an improved Hopcroft-Karp (HK) algorithm is presented to solve the problem of channel allocation, which has low complexity.

3.1 Power Allocation Problem

To simplify the power allocation problem, we assume that D2D pair j chooses to reuse the uplink channels of the CUs, then the transmission powers of D2D pair and CU are mainly considered in P2, which can be written as

$$P2 : \max_{\mathbf{P}^d, \mathbf{P}^c} \sum_{i=1}^m \sum_{j=1}^n R_{ij}^{ul,d} \quad (18)$$

s.t. (5), (6), (11), (12).

According to (18) and constraint (12), the problem P2 can be rewritten as

$$L(\{P_i^c, P_{ij}^d, \lambda\}) = \sum_{i=1}^m \sum_{j=1}^n \log\left(1 + \frac{P_{ij}^d \cdot h_{jj}}{P_i^c \cdot h_{i,j} + N_0}\right) - \lambda \left(\sum_{i=1}^m \sum_{j=1}^n P_{ij}^d - P_{total}^d\right), \quad (19)$$

where λ is the multiplication factor, and the Lagrangian dual function can be expressed as

$$G(\lambda) = \max_{P_{ij}^d, P_i^c} L(\{P_i^c, P_{ij}^d\}, \lambda) = \sum_{j \in N} G_j(\lambda) + \lambda P_{total}^d, \quad (20)$$

where $G_j(\lambda)$ can be written as

$$G_j(\lambda) = \max_{P_{ij}^d, P_i^c} \sum_{i \in M} \left[\log\left(1 + \frac{P_{ij}^d \cdot h_{jj}}{N_0 + P_i^c \cdot h_{ij}}\right) - \lambda P_{ij}^d \right], \quad (21)$$

We denote by $G_i^j(\lambda)$ the value when the D2D pair j reuses the uplink channel of CU i . Then the original optimization function can be converted into the following expression

$$G_i^j(\lambda) = \max_{P_{ij}^d, P_i^c} \sum_{j \in N} \left[\log\left(1 + \frac{P_{ij}^d \cdot h_{jj}}{N_0 + P_i^c \cdot h_{ij}}\right) - \lambda P_{ij}^d \right]. \quad (22)$$

Due to P_i^c and P_{ij}^d are coupled in the above problem, we decompose it into two levels [15]. Then we can get following formula according to (5) and (11)

$$\frac{\gamma_{th}^c}{h_{iB}} (P_{ij}^d \cdot h_{jB} + N_0) \leq P_i^c \leq P_i^{max}. \quad (23)$$

Thus, the optimal value of P_i^c is

$$P_i^* = \frac{\gamma_{th}^c}{h_{iB}} (P_{ij}^d \cdot h_{jB} + N_0) \quad (24)$$

For a given P_{ij}^d , (22) is a monotonically decreasing function of P_i^c . It shows that (22) can reach the maximum value when P_i^c obtain the minimum value. Then according to (24), (22) can be written as

$$G_i^j(\lambda) = \max_{P_{ij}^d} [\log(1 + \frac{P_{ij}^d \cdot h_{jj}}{P_{ij}^d \frac{\gamma_{ih}^c h_{jB} h_{ij}}{h_{iB}} + N_0 (1 + \frac{\gamma_{ih}^c h_{ij}}{h_{iB}})} - \lambda P_{ij}^d)] \quad (25)$$

By taking the second-derivative of (22), the value of the second-derivative is always less than 0. Thus, it is a convex function about P_{ij}^d . We take the first-derivative of formula $G_i^j(\lambda)$ for P_{ij}^d and set the value of first-derivative to 0. The optimal power P_{ij}^* can be obtained by the subgradient method [16], then the optimal power P_i^* can be obtained according to (24) and the value of P_{ij}^* .

3.2 Channel Allocation Problem

After channel assignment, we will get the optimal powers of D2D pairs and CUs, and then we specify some channel for each D2D pair. For CU i , if no D2D pair reuses its channel, the transmission rate of CU i on UL is as follows:

$$R_i^{ul,max} = \log(1 + \frac{P_i^{max} h_{iB}}{N_0}). \quad (26)$$

If the uplink channel of CU i is reused by the D2D pair j , the throughput gain of the system in UL communication phase is:

$$T_{ij}^{ul} = R_i^{ul,c} + R_{ij}^{ul,d} - R_i^{ul,max}, \quad (27)$$

Similarly, in DL communication phase, the throughput gain of system can be calculated. Then the optimal D2D-CUs problem turns to be a maximum weight bipartite matching problem, and it can be given by

$$P3 : \max_{\mathbf{a}^u, \mathbf{a}^d} \sum_{i=1}^m \sum_{j=1}^n a_{ij}^u T_{ij}^{ul} + a_{ij}^d T_{ij}^{dl} \quad (28)$$

$$s.t. (1), (2), (3), (4).$$

We represent Set \mathcal{D} as one group vertices of D2Ds pairs and Set \mathcal{C} as the other group vertices of channels which includes both the uplink and downlink subcarriers. Assuming the uplink channel of CU i is reused by D2D j , T_{ij}^{ul} represents the weights of the edge connecting D2D pair j and CU i . Similarly, if D2D pair j reuses the downlink channel of CU i , the weights of the edge can represent as T_{ij}^{dl} . In addition, to guarantee that each D2D pair can reuse the same number of channels, we use q_j to represent the maximum number of channels for D2D j and set the value of q_j is equals to 3. Then we propose an a low-complexity algorithm to select channels for each D2D pair, and we can get the result of channel matching in Algorithm 1 based on the Hopcroft-Karp algorithm [17].

Algorithm 1. an improved HK algorithm for channel allocation

```

1: Initial :  $|L_{ij}| = 0$ ,  $\mathbf{a} = [a_{ij}^u, \dots, a_{ij}^d] = \mathbf{0}$ ,
2: Input : the optimal powers  $\mathbf{P}_i^*$  and  $\mathbf{P}_{ij}^*$ 
3: Output :  $\mathbf{a} = [a_{ij}^u, \dots, a_{ij}^d]$ 
4: while  $|L_{ij}| \leq q_j$  and  $\mathcal{D}$  is not empty do
5:   /*channel selection
6:   we first choose D2D pair  $j$  from  $\mathcal{D}$ , and execute HK algorithm to select channel
   for D2D pair, then add the matching channel to  $L_{ij}$ 
7:   if  $L_{ij}$  coexists  $L_{ij}^d$  and  $L_{ij}^u$  then
8:     /* exist channel conflict according to (4)
9:     not remove D2D  $j$  from  $\mathcal{D}$  and not remove CU  $i$  from  $\mathcal{C}$ 
10:  else if  $L_{ij}$  only exist  $L_{ij}^u$  then
11:    set  $a_{ij}^u = 1$  and remove the UL of CU  $i$  from  $\mathcal{C}$ 
12:  else
13:    set  $a_{ij}^d = 1$  and remove the DL of CU  $i$  from  $\mathcal{C}$ 
14:  end if
15:  if  $L_{ij}$  is a maximum match then
16:    remove D2D  $j$  from  $\mathcal{D}$  and remove the select channel of D2D  $j$  from  $\mathcal{C}$ 
17:  else
18:    continue
19:  end if
20: end while

```

4 Performance Evaluation

4.1 Experiment Parameter

In the simulation experiment, we consider the single-cell scenario, where BS is deployed in the center of the cell, CU and D2D are uniformly distributed, and the radius of cell is 500 m and 800 m, respectively. The specific parameter settings are shown in Table 1. Then, we compare the resource allocation algorithm (JAUD)

Table 1. Simulation parameters

Parameter	Value
Bandwidth	0.5 MHz
Noise spectral density	-144 dBm
Path loss exponent	3
Maximum transmission power of CU	20 dBm
Maximum transmission power of D2D	22 dBm
SINR threshold of CU and D2D	7 dB
D2D cluster radius	5-30 m
Number of CUEs (M) and D2D pair (N)	50, 10

proposed in this paper based on the joint reuse of uplink and downlink spectrum with the following basic schemes:

- all D2D links only reuse the uplink resources;
- all D2D links only reuse downlink resources;
- the Hungarian algorithm in uplink and downlink resource allocation [12].

4.2 Comparative Analysis of Experiments

Figure 2 compares the performance of different schemes against different numbers of CUs. It is evident that D2D throughput increases with the number of CUs because D2D pair can reuse more resources. In JUAD scheme, since each D2D pair can reuses channels of multiple CUs, the system performance is better compared with others. Moreover, in OU scheme, due to the impact of interference from CUs on D2D pair, the performance of OU scheme is better than OD scheme.

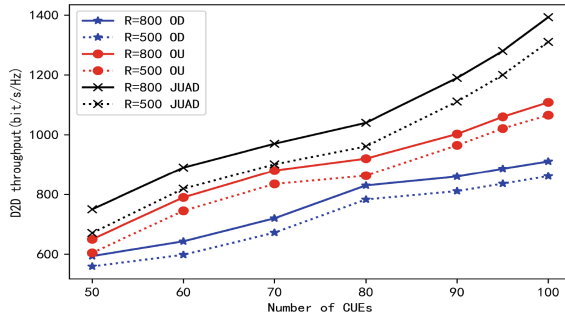


Fig. 2. D2D throughput with different numbers of cellular users (CUs)

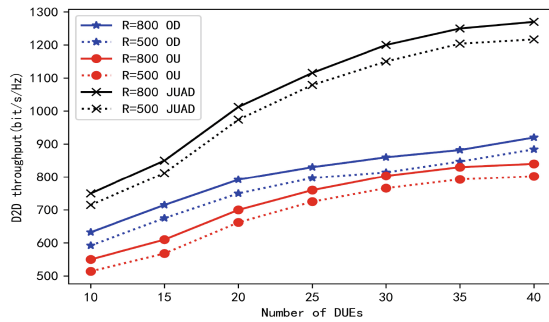


Fig. 3. D2D throughput versus the number of D2D pairs.

Figure 3 illustrates the impact of the number of D2D pairs on the performance of D2D networks. Obviously, the performance of D2D networks increases with the number of D2D pairs. In JUAD scheme, the throughput increases rapidly, but the performance of D2D networks increases slowly in others scheme, this is because each resource of CU can only be reused by at most one D2D pair.

Figure 4 shows the influence of different D2D cluster radius on the total throughput of D2D networks under three schemes. It can be seen that the increasement of distance in D2D cluster will result in the decrease of D2D throughput.

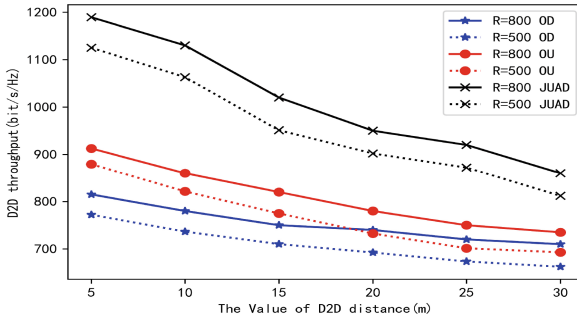


Fig. 4. D2D throughput versus D2D cluster radius.

In Fig. 5, we note that the throughput of D2D networks with radius $R = 800$ m is obviously better than that with radius $R = 500$ m in JUAD scheme. It is owing to the interference gain declines as the cell radius increases, so D2D pair has more chances to select some good channels.

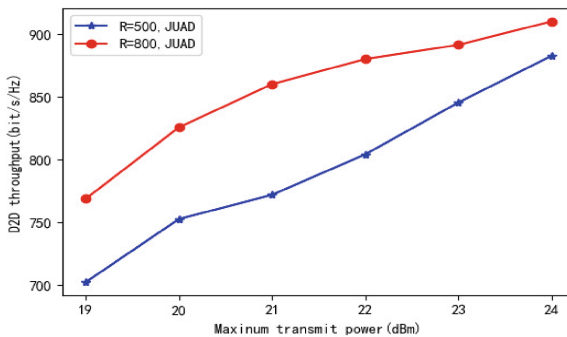


Fig. 5. D2D throughput versus transmission power.

Figure 6 illustrates the access rate versus the number of D2D pairs. With the increasement of D2D pair, the access rate of D2D pairs monotonically decreases,

which means that when the number of CUs is fixed, the number of D2D pairs sharing the channels of CUs are limited. However, compared with the OU and OD scheme, the access rate of D2D pair in JUAD scheme decreases slowly. That is because D2D pair can joint reuse uplink and downlink channels of CUs in the JUAD scheme.

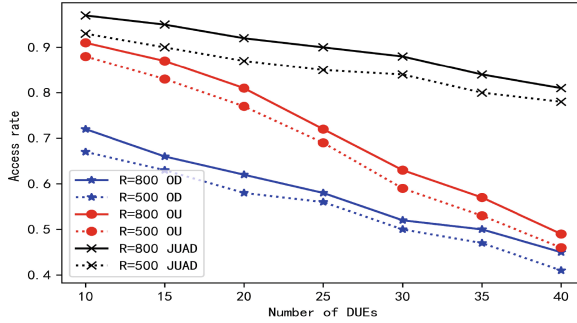


Fig. 6. The D2D access rate versus the number of D2D users.

Figure 7 shows the influence of algorithm complexity with different system users. It can be seen that algorithm running time increases with the increase of system users, in which the running time of OU, OD and JAUD scheme are close. This is because both OU and OD are based on the Hopcroft-Karp algorithm. Since JAUD algorithm carried out mode selection, the running time of JAUD algorithm is longer than OU and OD scheme.

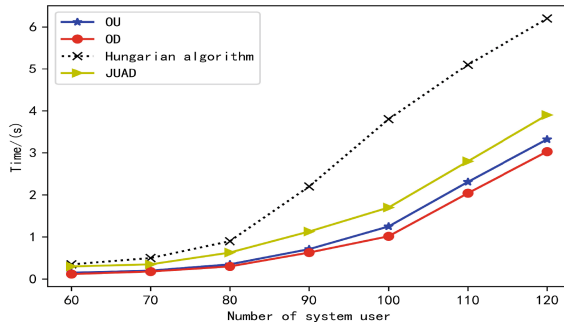


Fig. 7. The running times of different algorithms

5 Conclusions

In this paper, we investigate the joint uplink and downlink resource allocation problem in D2D communication networks, where each D2D pair can reuse the

resources of multiple CUs. To maximize the overall throughput of D2D networks, we model the JUAD resource allocation problem as a mixed integer nonlinear programming problem, which jointly consider the influence of channel selection and transmit power value. Afterwards, we decouple the problem into two sub-problems, namely, power allocation and channel allocation. Then, the power allocation problem is settled with convex optimization techniques to get the optimum value of power. Next, we propose an improved Hopcroft-Karp algorithm to solve the problem of channel allocation, which has a lower time complexity compared with traditional channel allocation approaches. Simulation results show that our proposed approach achieves a near-optimal solution to the JUAD problem.

Acknowledgment. This work is supported by the Natural Science Foundation of China (No. 61872104), the Natural Science Foundation of Heilongjiang Province in China (No. F2016009), the Fundamental Research Fund for the Central Universities in China (No. HEUCF180602) and the Tianjin Key Laboratory of Advanced Networking (TANK), College of Intelligence and Computing, Tianjin University, Tianjin China, 300350.



References

1. Torre, R., Fitzek, F.H.P.: A study on data dissemination techniques in heterogeneous cellular networks. In: Sucasas, V., Mantas, G., Althunibat, S. (eds.) BROADNETS 2018. LNICST, vol. 263, pp. 169–179. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05195-2_17
2. Wu, D., Wang, J., Hu, R.Q., Cai, Y., Zhou, L.: Energy-efficient resource sharing for mobile device-to-device multimedia communications. *IEEE Trans. Veh. Technol.* **63**(5), 2093–2103 (2014)
3. Han, M.-H., Kim, B.-G., Lee, J.-W.: Subchannel and transmission mode scheduling for D2D communication in OFDMA networks. In: 2012 IEEE Vehicular Technology Conference (VTC Fall), pp. 1–5. IEEE (2012)
4. Sikora, M., Laneman, J.N., Haenggi, M., Costello Jr., D.J., Fuja, T.E.: On the optimum number of hops in linear wireless networks. In: Proceedings of IEEE Information Theory Workshop, pp. 165–169 (2004)
5. Akyildiz, I.F., Lee, W.Y., Vuran, M.C., Mohanty, S.: Next generation/dynamic spectrum access/cognitive radio wireless networks: a survey. *Comput. Netw.* **50**, 2127–2159 (2006)
6. Menon, R., Buehrer, R.M., Reed, J.H.: Outage probability based comparison of underlay and overlay spectrum sharing techniques. In: Proceedings of IEEE DySPAN, vol. 5, pp. 101–109 (2005)
7. Peha, J.M.: Approaches to spectrum sharing. *IEEE Commun. Mag.* **43**(2), 10–12 (2005)
8. Janis, P., Koivunen, V., Ribeiro, C., Korhonen, J., Doppler, K., Hugl, K.: Interference-aware resource allocation for device-to-device radio underlaying cellular networks. In: VTC Spring 2009-IEEE 69th Vehicular Technology Conference, pp. 1–5. IEEE (2009)
9. Ma, C., Liu, J., Tian, X., Hui, Y., Cui, Y., Wang, X.: Interference exploitation in D2D-enabled cellular networks: a secrecy perspective. *IEEE Trans. Commun.* **63**(1), 229–242 (2015)

10. Xiao, S., et al.: Joint uplink and downlink resource allocation in full-duplex OFDMA networks. In: 2016 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2016)
11. Zhao, P., Yu, P., Feng, L., Li, W., Qiu, X.: Gain-aware joint uplink-downlink resource allocation for device-to-device communications. In: 2017 IEEE 85th Vehicular Technology Conference (VTC Spring), pp. 1–5. IEEE (2017)
12. Song, X., Han, X., Ni, Y., Dong, L., Qin, L.: Joint uplink and downlink resource allocation for D2D communications system. *Future Internet* **11**(1), 12 (2019)
13. Kai, C., Xu, L., Zhang, J., Peng, M.: Joint uplink and downlink resource allocation for D2D communication underlying cellular networks. In: 2018 10th International Conference on Wireless Communications and Signal Processing (WCSP), pp. 1–6. IEEE (2018)
14. Malandrino, F., Limani, Z., Casetti, C., Chiasserini, C.-F.: Interference-aware downlink and uplink resource allocation in hetnets with D2D support. *IEEE Trans. Wirel. Commun.* **14**(5), 2729–2741 (2015)
15. Sasao, T., Matsuura, M.: A method to decompose multiple-output logic functions. In: Proceedings of the 41st Annual Design Automation Conference, pp. 428–433. ACM (2004)
16. Kiwiel, K.C.: An aggregate subgradient method for nonsmooth convex minimization. *Math. Program.* **27**(3), 320–341 (1983)
17. Gabow, H.N.: Scaling algorithms for network problems. In: 24th Annual Symposium on Foundations of Computer Science (SFCS 1983), pp. 248–258. IEEE (1983)



FaLQE: Fluctuation Adaptive Link Quality Estimator for Wireless Sensor Networks

Wei Liu¹ , Yu Xia¹ , and Rong Luo² 

¹ School of Electrical and Electronic Engineering,
Chongqing University of Technology, Chongqing 400054, China
liu-wei@cqut.edu.cn

² Department of Electronic Engineering, Tsinghua University,
Beijing 100084, China

Abstract. Accurate link quality estimation is a prerequisite for efficient routing in wireless sensor networks. Good link quality estimators should provide agility and stability simultaneously, which not only filter out transient link quality fluctuations but also respond quickly when sudden changes arise. However, only stability or agility is considered as optimization goal in existing estimators, so their performance is always below expectations. In this paper, a fluctuation adaptive link quality estimator is proposed, which adjusts smoothing factor of the estimation dynamically according to the degree of link quality fluctuations and achieves equilibrium of stability and agility. Experimental results show that stability of the proposed estimator is same as that of existing stable estimators when there are transient fluctuations, and agility of the proposed estimator is same as that of existing agile estimators when sudden changes arise. More importantly, compared with existing estimators, the estimate error of the proposed one is reduced by 22.5%–31.8% for different link characteristics.

Keywords: Link quality estimation · Wireless sensor networks · Fluctuation adaptive · Stability · Agility

1 Introduction

Wireless sensor networks (WSNs) are multi-hop self-organizing networks composed of hundreds and thousands of sensor nodes with parameter sensing, information processing and wireless communication capabilities. They have been successfully used in many fields including military investigation, environmental monitoring, industrial control and medical care. WSNs typically use low power RF transceivers, which make the wireless links less stable and have many fluctuations. In order to find the best end-to-end routes and improve transmission efficiency, real-time accurate link quality estimation is necessary. As a result, a good link quality estimator (LQE) is of the essence for the design of these networks.

Evaluation parameters of LQEs mainly include accuracy, agility, stability and overhead. Good LQEs should provide agility and stability simultaneously, which not only filter out transient link quality fluctuations but also respond quickly when sudden changes arise. However, from the traditional point of view, stability and agility are at

odds [1]. Improving stability will always be at the cost of agility and vice versa. As a result, only stability or agility is considered as optimization goal in existing estimators [2–4], so their performance is always below expectations. It is due to the fact that existing LQEs are lack of effective perception and self-adaption to link dynamics. In this paper, a Fluctuation adaptive Link Quality Estimator (FaLQE) is proposed, which adjusts smoothing factor of the estimation according to changing degree of packet reception rate (PRR) of adjacent estimation windows dynamically and achieves equilibrium of stability and agility accordingly.

The rest of this paper is organized as follows. In Sect. 2, related works in link quality estimation are given. This is followed by experimental setup and analysis in Sect. 3. Design motivation and algorithm description of the proposed estimator are described in Sect. 4. Performance comparisons with other estimators are discussed in Sect. 5. Finally, conclusions are presented and suggestions are made for future works.

2 Related Works

Existing LQEs can be classified into four categories: hardware-based [5–8], software-based [2, 9], hybrid metrics-based [3, 4, 10–12] and machine learning-based [13–15]. Hardware-based estimators predict link quality by directly using physical layer parameters such as received signal strength indicator (RSSI) and link quality indicator (LQI) or based on relationships between these parameters and PRR. Although hardware-based estimators have advantages of low overhead and high sensitivity, they may overestimate the link quality by ignoring information from lost data packets, because physical layer parameters can only be obtained from successfully received data packets.

Software-based estimators predict link quality by computing the packet reception rate or the number of transmissions over a defined period of time, which is usually called a window. ETX (Expected Transmission Count) is a software-based estimator, which is obtained by computing the inverse of the product of PRR of the forward link and PRR of the backward link [9]. WMEWMA (Window Mean with Exponentially Weighted Moving Average) is another software-based estimator, which uses an exponentially weighted moving average filter to process the window means of PRR [2]. It solves the problem that raw PRR is too agile. However, when there are sudden changes in the link, it cannot keep up with the changes as quickly as possible. Performance of the software-based LQEs depends on the choices of window size: with smaller windows, the estimator is more agile but may cause unnecessary switching of routes; with larger windows, the estimator is more stable but cannot respond quickly when sudden changes arise [1].

Hybrid metrics-based estimators combine multiple metrics together to estimate link quality. Four-Bit uses LQI to quickly identify whether the link has high quality or not, and then estimate the link quality through calculating uplink and downlink’s ETX [10]. EasiLQE dynamically selects the size of next detection window by average RSSI, and outputs estimated PRR of current detection window through an error-based filter [3]. As the smoothing factor of the error-based filter changes too fast when the link fluctuates, EasiLQE cannot filter out transient link fluctuations effectively. F-LQE (Fuzzy

Link Quality Estimator) processes four metrics of the link based on fuzzy logic [11]. As F-LQE is too stable, some researchers attempt to achieve more agile and accurate estimations by adjusting metrics of fuzzy logic [4, 12]. ELQET (Enhanced Link Quality Estimation Technique) uses four link metrics, PRR, stability attribute, average signal-to-noise ratio (SNR), and average LQI to describe a link, and estimate the link quality by fuzzy logic and exponentially weighted moving average filters [4]. Although fuzzy logic-based LQEs are very stable, their agility is relatively poor due to the requirement for combining multiple link metrics.

In recent years, machine learning-based estimators attract the attention of researchers, which are used to improve agility and accuracy of link quality estimation. However, there are some defects for this kind of estimator. On one hand, machine learning algorithms are too complicated and difficult to be executed efficiently on sensor nodes with limited computing power and memory. On the other hand, one intention of link quality estimation is to reduce energy overhead of network transmissions. However, excessive energy overhead brought by executing machine learning algorithms is obviously contrary to this intention.

In summary, only stability or agility is considered as optimization goal in existing LQEs, so their performance is always below expectations. It is due to the fact that existing LQEs are lack of effective perception and self-adaption to link dynamics. The contribution of this paper is that a fluctuation adaptive link quality estimator is proposed, which adjusts smoothing factor of the estimation dynamically according to the degree of link quality fluctuations and achieves equilibrium of stability and agility.

3 Experimental Setup and Analysis

In order to obtain wireless links with different characteristics, three different experimental fields are chosen, as shown in see Fig. 1. Among these fields, playground is a typical outdoor environment which has simple propagation channel and low external interferences. On the contrary, corridor of our building is a typical indoor environment which has complex propagation channel and high external interferences. Experiments are conducted using two TelosB nodes, one as transmitter and the other as receiver. TelosB uses TinyOS 2.1 operating system and is programmed with NesC language [16]. Changing the distance between transmitter and receiver to produce different link qualities. 500 packets are sent in each experiment, and PRR is calculated using the number of successfully received data packets.

Wireless links are typically classified into three categories according to different PRR values, which are good link ($PRR > 80\%$), moderate link ($20\% \leq PRR \leq 80\%$) and bad link ($PRR < 20\%$) [17]. This classification evaluates link quality from a long-term statistical perspective, which ignores inherent transient fluctuations of wireless links. In order to analyze PRR distribution under different link qualities, 50 groups of data are chosen randomly from links with different qualities. Then, distributions of PRR are calculated by taking windows with different sizes. Figures 2 and 3 show the cumulative distribution functions (CDF) of PRR under good, moderate and bad links

for window size of 0.5 s and 1.25 s respectively. It can be seen that cumulative probability for $PRR > 80\%$ under good link is greater than 0.95, cumulative probability for $PRR < 20\%$ under bad link is greater than 0.85, and cumulative probability for $20\% \leq PRR \leq 80\%$ under moderate link exceeds 0.6. It is shown that using statistical means of PRR to describe link qualities is reasonable to some extent. However, transient values of PRR under good link may be less than 0.2, and transient values of PRR under bad link may also be greater than 0.8. It's same for moderate link. Moreover, if transient values of PRR are changed from 0.8 to 0.2 suddenly, it has large probability to change from good link to bad link, and vice versa. This indicates that inherent transient fluctuations of wireless links are ignored unwisely when using statistical means of PRR to describe link qualities.

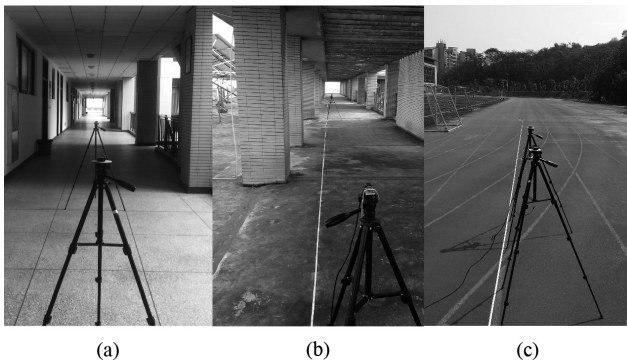


Fig. 1. Environments for link quality experiments: (a) Corridor, (b) Rooftop, and (c) Playground.

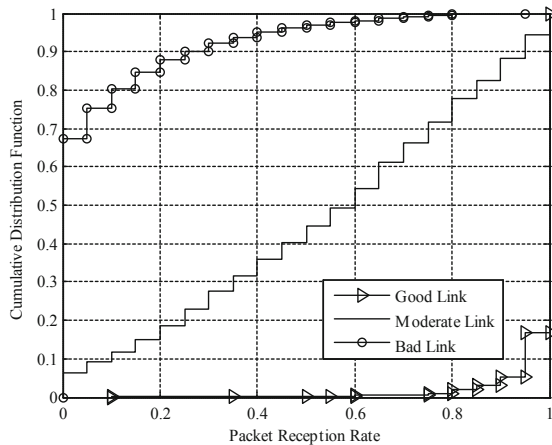


Fig. 2. CDFs of PRR under different links when window size is 0.5 s.

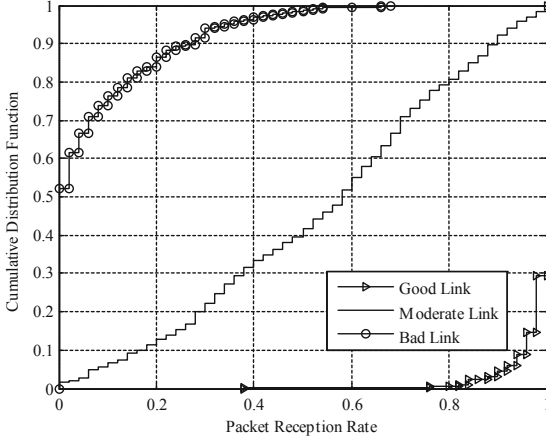


Fig. 3. CDFs of PRR under different links when window size is 1.25 s

4 Fluctuation Adaptive Link Quality Estimator

From the above analysis, it is obvious that wireless link is mutable in practical environments. The PRR may change from 0.8 to 0.2 in short time and vice versa. In other words, it is not enough to only rely on the long-term statistical means of PRR to describe link qualities. It is necessary to take transient fluctuations of PRR into consideration carefully. Therefore, difference between PRR of two consecutive time windows is used as an indicator to determine the degree of transient fluctuations of the link. We define fluctuation coefficient $\Delta(i)$ as follows:

$$\Delta(i) = |PRR(i) - PRR(i - 1)| \quad (1)$$

where i denotes the i -th window.

Different fluctuation coefficients correspond to different degrees of link fluctuations. According to the values of $\Delta(i)$, links can be classified into three categories:

Definition 1 (Stable link). $\Delta(i) < 0.2$, indicating that transient fluctuations of the link are small and negligible;

Definition 2 (Fluctuating link). $0.2 \leq \Delta(i) \leq 0.5$, indicating that the link quality is unstable and there are nonnegligible transient fluctuations;

Definition 3 (Sudden changed link). $\Delta(i) > 0.5$, indicating that there are large fluctuations in the link, and the probability that a permanent change in link quality would occur is high.

The LQE should be self-adaptive to different degrees of link fluctuations: it should not only filter out transient link quality fluctuations but also respond quickly when sudden changes arise. This requires the estimator to be able to effectively sense link fluctuations and make corresponding adjustments. Unfortunately, existing LQEs often ignore this point, which makes their performance always below expectations. By using

fluctuation coefficient previously defined, we proposed FaLQE, a fluctuation adaptive link quality estimator. FaLQE is a software-based estimator, in which the estimated PRR of the i -th window is:

$$PRR_{FaLQE}(i) = \alpha(i) \times PRR_{FaLQE}(i - 1) + (1 - \alpha(i)) \times PRR(i) \quad (2)$$

where $\alpha(i)$ denotes fluctuation adaptive smoothing factor, which is defined as follows:

$$\alpha(i) = \begin{cases} 0.6 & \Delta(i) < 0.2 \\ 0.4 & 0.2 \leq \Delta(i) \leq 0.5 \\ 0.1 & \Delta(i) > 0.5 \end{cases} \quad (3)$$

where i denotes the i -th window. The values of $\alpha(i)$ are chosen with the following considerations:

- (1) When transient fluctuations of the link are small and negligible, which means a stable link, the LQE should filter out these fluctuations effectively, so a larger smoothing factor is needed;
- (2) When the probability that a permanent change in link quality would occur is high, which means a sudden changed link, the LQE should keep up with the change as quickly as possible, so a smaller smoothing factor is needed;
- (3) The smoothing factor under a fluctuating link should be between above two cases.

FaLQE chooses corresponding smoothing factors according to different fluctuation coefficients in order to achieve self-adaption to link dynamics. The algorithm flow chart of FaLQE is shown in Fig. 4.

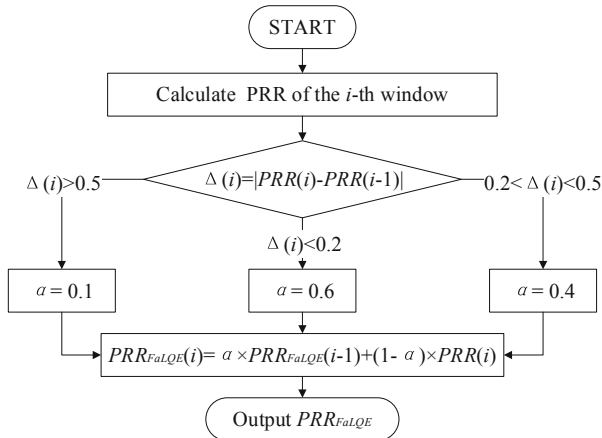


Fig. 4. The algorithm flow chart of FaLQE.

5 Performance Comparison

To demonstrate the performance benefits of FaLQE, three popular LQEs, namely WMEWMA [2], EasiLQE [3] and ELQET [4] are chosen. Stability and agility of these estimators are compared. WMEWMA is a software-based estimator with high stability as its optimization goal. The estimated PRR of the i -th window is:

$$PRR_{WMEWMA}(i) = \alpha \times PRR_{WMEWMA}(i-1) + (1 - \alpha) \times PRR(i) \quad (4)$$

where the smoothing factor $\alpha = 0.6$, and $PRR(i)$ is the packet reception rate of the i -th window.

EasiLQE is a hybrid metrics-based estimator with high agility as its optimization goal, in which RSSI is used to assist PRR estimation. The estimated PRR of the i -th window is:

$$PRR_{EasiLQE}(i) = \alpha_t \times PRR_{EasiLQE}(i) + (1 - \alpha_t) \times PRR(i) \quad (5)$$

where α_t is the smoothing factor and calculated as follows:

$$\alpha_t = 1 - \left(\frac{\Delta_t}{\Delta_{max}} \right) \quad (6)$$

where Δ_t denotes estimation error and Δ_{max} is the maximum of m consecutive Δ_t values.

ELQET is also a hybrid metrics-based estimator, but its optimization goal is high stability. The estimated value of the i -th window is:

$$ELQET(i) = \lambda \times ELQET(i-1) + (1 - \lambda) \times \mu(i) \quad (7)$$

where $\lambda = 0.8$. Four metrics, namely PRR, average SNR (ASNR), average LQI (ALQI), and stability attribute (SA) are combined using fuzzy logic to compute $\mu(i)$:

$$\begin{aligned} \mu(i) &= \alpha \min(\mu_{PRR}(i), \mu_{SA}(i), \mu_{ASNR}(i), \mu_{ALQI}(i)) \\ &+ (1 - \alpha) \text{mean}(\mu_{PRR}(i), \mu_{SA}(i), \mu_{ASNR}(i), \mu_{ALQI}(i)) \end{aligned} \quad (8)$$

where $\alpha = 0.6$.

In order to obtain adequate data for performance comparisons of these estimators, test was conducted in the corridor for a long time. Multiple link conditions such as stable links, fluctuating links and sudden changed links are recorded. Three links with 500 s each are chosen and PRR is calculated every 50 s, as shown in Fig. 5. The characteristics and objectives of these links are as follows:

- (1) Link 1: Transient fluctuations of this link is small and negligible, which conforms to the definition of stable link. It is mainly used to evaluate stability of each estimator.
- (2) Link 2: The quality of this link is unstable and there are nonnegligible transient fluctuations, which conforms to the definition of fluctuating link. It is used to evaluate both stability and agility of each estimator.

- (3) Link 3: A change from good link to bad link occurs in this link, which conforms to the definition of sudden changed link. It is mainly used to evaluate agility of each estimator.

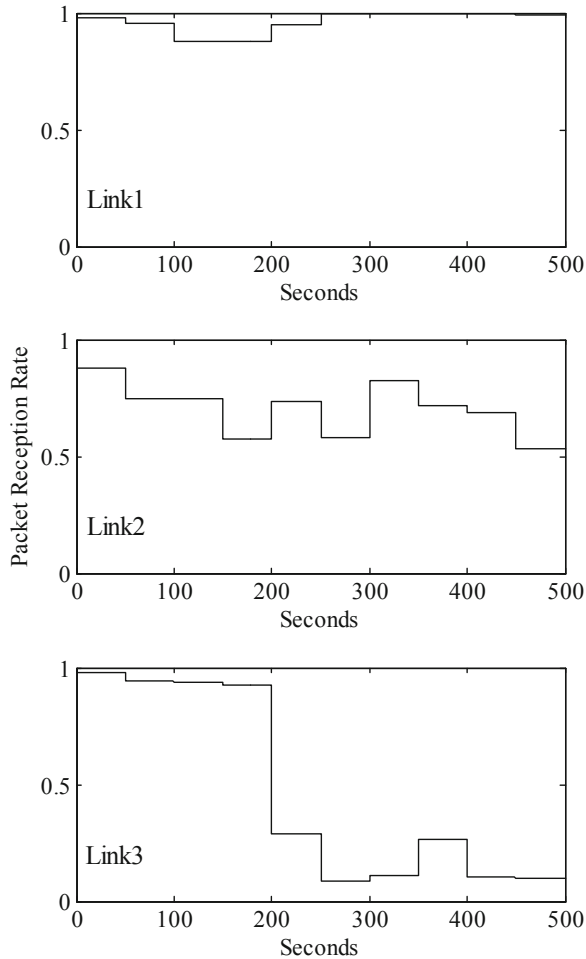


Fig. 5. Practical PRR of link 1, link 2 and link 3.

Figure 6 shows the response curves of FaLQE, WMEWMA, ELQET and EasiLQE under three links described above. It is obvious that for link 1 with small transient fluctuations, stability of FaLQE is same as WMEWMA and ELQET, namely stable estimators. As EasiLQE is optimized for high agility, its estimations present larger fluctuations and cannot filter out small transient fluctuations. In addition, FaLQE can keep track of link changes quickly, which is better than WMEWMA and ELQET.

For link 2 with larger transient fluctuations, the estimated values of EasiLQE are more fluctuating with maximum estimate error close to 0.5. In contrast, FaLQE shows better performance, it not only filters out unnecessary fluctuations but also keeps track of link changes.

For link 3 with a sudden change, WMEWMA and ELQET both take long time to keep track of the change. Instead, FaLQE can respond quickly within one window. Although EasiLQE can also respond quickly, its estimations fluctuate too much, even with estimate error more than 0.7.

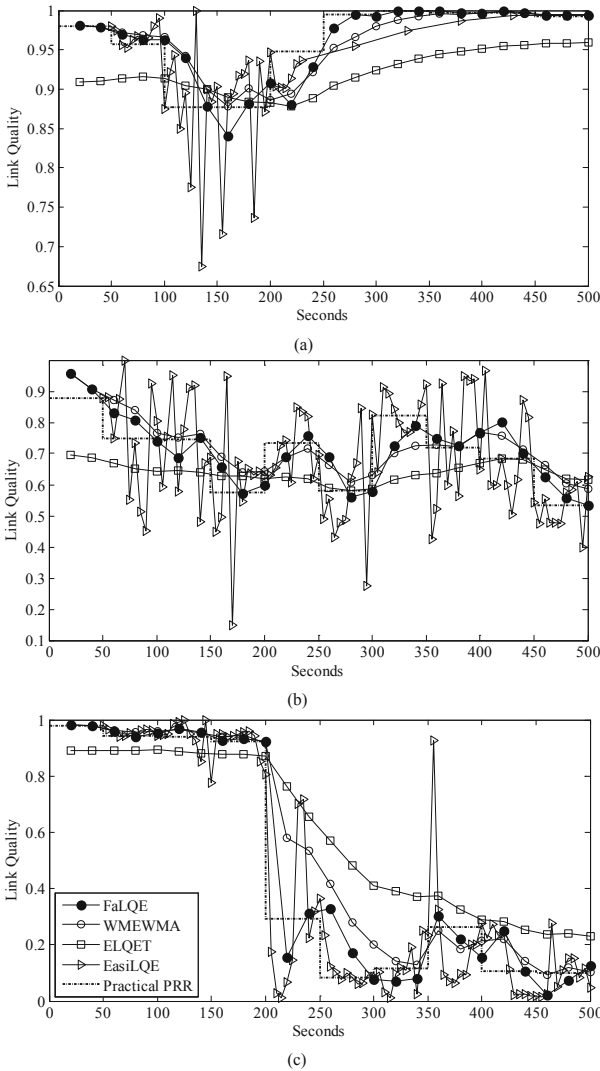


Fig. 6. Response curves of FaLQE, WMEWMA, ELQET and EasiLQE: (a) link 1,(b) link 2, (c) link 3.

From the above analysis, it can be concluded that the response curve of our proposed estimator is most consistent with practical PRR. For links with transient fluctuations, stability of FaLQE is same as that of existing stable estimators. Meanwhile, for links with sudden changes, agility of FaLQE is same as that of existing agile estimators. Estimated value's coefficient of variation (CV) is generally used in the literature to assess stability and agility of LQEs quantitatively, which is defined as the ratio of the standard deviation to the mean of estimated values [1]:

$$CV = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^n \left(E(i) - \frac{1}{n} \sum_{i=1}^n E(i) \right)^2}}{\frac{1}{n} \sum_{i=1}^n E(i)} \quad (9)$$

where $E(i)$ denotes the estimated value of the i -th window and n is the number of estimated values. For stable and fluctuating links, smaller CV means more stable estimations. For sudden changed links, larger CV means more agile estimations.

Figure 7 shows coefficients of variation for FaLQE, WMEWMA, ELQET and EasiLQE under three links described above. For link 1 and link 2, the CV of FaLQE is close to that of WMEWMA and a little higher than that of ELQET, but significantly lower than the CV of agile estimator EasiLQE. For link 3, the CV of FaLQE is lower than that of EasiLQE, but higher than the CVs of WMEWMA and ELQET. Consequently, we can conclude that FaLQE provides agility and stability simultaneously, which not only filter out transient link quality fluctuations but also respond quickly when sudden changes arise.

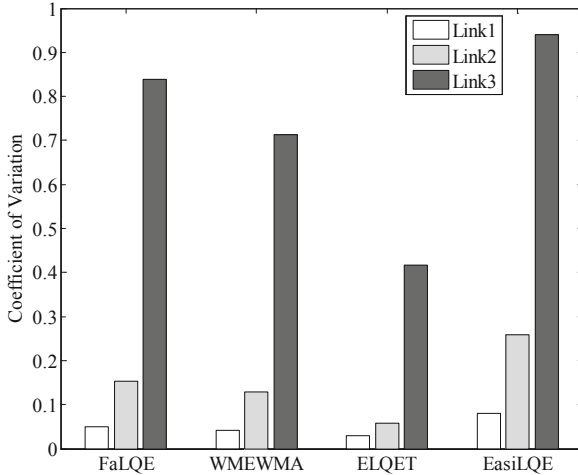


Fig. 7. Coefficients of variation for LQEs under different links.

Figure 8 shows the root mean square errors (RMSE) between estimated values and practical PRR. The smaller the RMSE is, the more accurate the LQE is. It can be concluded that RMSE of FaLQE is the smallest under all kinds of links, which indicates that FaLQE adapts to link dynamics better and obtains the closest estimation to practical PRR. Compared with WMEWMA, ELQET and EasiLQE, the estimate error of FaLQE is reduced by 22.5%–31.8%.

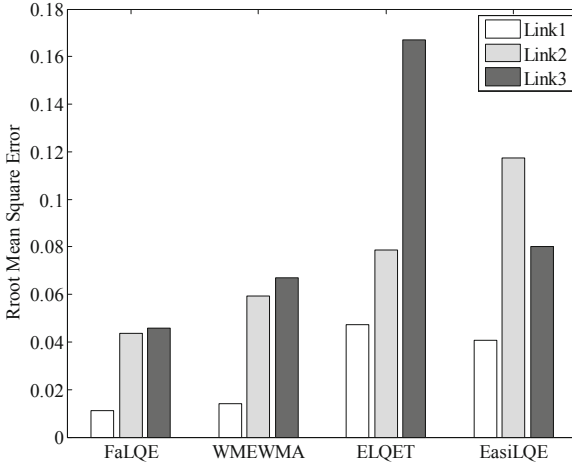


Fig. 8. RMSEs for LQEs under different links.

6 Conclusions and Future Works

Links in wireless sensor networks are mutable in practical environments. However, existing LQEs are lack of effective perception and self-adaption to link dynamics. As a result, they cannot provide agility and stability simultaneously. In this paper, fluctuation coefficient is defined as an indicator to determine the degree of transient fluctuations of the link, and links are classified into three categories: stable link, fluctuating link and sudden changed link. Based on such a classification, a fluctuation adaptive link quality estimator FaLQE is proposed. In order to achieve self-adaption to link dynamics, the proposed FaLQE adjusts smoothing factor of the estimation according to the changing degree of packet reception rate of adjacent estimation windows dynamically.

Experimental results show that stability of the proposed estimator is same as that of existing stable estimators, such as WMEWMA and ELQET, when there are small transient fluctuations, and agility of the proposed estimator is same as that of existing agile estimators, such as EasiLQE, when sudden changes arise. More importantly, due to equilibrium of stability and agility, the estimate error of FaLQE is reduced by 22.5%–31.8% for different link characteristics compared with other estimators. In future works, comprehensive tests and in-depth analysis will be conducted to evaluate performance of FaLQE under different kinds of links, including its accuracy, agility, stability and

overhead. At the same time, FaLQE will be integrated into existing protocol stacks, for assessing its impact on upper layer protocols and network performance.

Acknowledgments. This work is supported in part by National Natural Science Foundation of China (Grant No. 61601069), Chongqing Research Program of Basic Research and Frontier Technology (Grant No. cstc2017jcyjAX0254), and Scientific and Technological Research Program of Chongqing Municipal Education Commission (Grant No. KJ1600935).

References

1. Baccour, N., et al.: Radio link quality estimation in wireless sensor networks: a survey. *ACM Trans. Sens. Netw.* **8**(4), 1–33 (2012)
2. Woo A, and Culler D.: Evaluation of efficient link reliability estimators for low-power wireless networks. Technical report, UCB/CSD-03-1270, EECS Department, University of California, Berkeley (2003). <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2003/CSD-03-1270.pdf>. Accessed 11 June 2019
3. Huang, T., Li, D., Zhang, Z., Cui, L.: Bursty-link-aware adaptive link quality estimation method. *J. Commun.* **33**(6), 30–39 (2012). (in Chinese)
4. Jayasri, T., Hemalatha, M.: Link quality estimation for adaptive data streaming in WSN. *Wirel. Pers. Commun.* **94**, 1543–1562 (2017)
5. Senel, M., Chintalapudi, K., Lal, D., Keshavarzian, A., Coyle, E.J.: A Kalman filter based link quality estimation scheme for wireless sensor networks. In: *IEEE Global Telecommunications Conference (GLOBECOM)*, pp. 875–880. IEEE, Washington D.C. (2007)
6. Shu, J., Tao, J., Liu, L., Chen, Y., Zang, C.: CCI-based link quality estimation mechanism for wireless sensor networks under non-perceived packet loss. *J. China Univ. Posts Telecommun.* **20**(1), 1–10 (2013)
7. Boano, C.A., Zúñiga, M.A., Voigt, T., Willig, A., Romer, K.: The triangle metric: fast link quality estimation for mobile wireless sensor networks. In: *International Conference on Computer Communications and Networks (ICCCN)*, pp. 1–7. IEEE, Zurich (2010)
8. Gomes, R.D., Queiroz, D.V., Filho, A.C.L., Fonseca, I.E., Alencar, M.S.: Real-time link quality estimation for industrial wireless sensor networks using dedicated nodes. *Ad Hoc Netw.* **59**, 116–133 (2017)
9. Couto, D.S.J.D., Aguayo, D., Bicket, J., Morris, R.: A high-throughput path metric for multi-hop wireless routing. In: *Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 134–146. ACM, San Diego (2003)
10. Fonseca, R., Gnawali, O., Jamieson, K., Levis, P.: Four bit wireless link estimation. In: *Proceedings of the International Workshop on Hot Topics in Networks (HotNets VI)*, pp. 1–6. ACM, Atlanta (2007)
11. Baccour, N., Koubâa, A., Youssef, H., Ben Jamâa, M., do Rosário, D., Alves, M., Becker, L. B.: F-LQE: a fuzzy link quality estimator for wireless sensor networks. In: Silva, J.S., Krishnamachari, B., Boavida, F. (eds.) *EWSN 2010*. LNCS, vol. 5970, pp. 240–255. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-11917-0_16
12. Rekić, S., Baccour, N., Jmaiel, M., Drira, K.: Low-power link quality estimation in smart grid environments. In: *Proceedings of International Wireless Communications and Mobile Computing Conference (IWCMC)*, pp. 1211–1216. IEEE, Dubrovnik (2015)
13. Liu, T., Cerpa, A.E.: Foresee (4C): wireless link prediction using link features. *Int. Conf. on Information Processing in Sensor Networks (IPSN)*, pp. 294–305. IEEE, Chicago (2011)

14. Liu, T., Cerpa, A.E.: TALENT: temporal adaptive link estimator with no training. In: ACM Conference on Embedded Network Sensor Systems (SenSys), pp. 253–266. ACM, Toronto (2012)
15. Marinca, D., Minet, P.: On-line learning and prediction of link quality in wireless sensor networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 1245–1251. IEEE, Austin (2014)
16. MEMSIC Inc.: TelosB datasheet. http://www.memsic.cn/userfiles/files/6020-0094-02_B_TELOSB.pdf. Accessed 1 June 2019
17. Srinivasan, K., Dutta, P., Tavakoli, A., Levis, P.: An empirical study of low-power wireless. ACM Trans. Sens. Netw. **6**(2), 1–49 (2010)



Rate Adaptive Broadcast in Internet of Things

Linghe Kong¹(✉), Zhe Wang¹, Yongshuai Duan¹, Tong Meng², Fan Wu¹,
and Guihai Chen¹

¹ Shanghai Jiao Tong University, Shanghai, China

{linghe.kong, wang-zhe, dys1998}@sjtu.edu.cn, {fwu, gchen}@cs.sjtu.edu.cn

² University of Illinois at Urbana-Champaign, Urbana, IL, USA
tongm2@illinois.edu

Abstract. This paper presents Rate Adaptive Broadcast (RAB), a novel wireless design that enables the rate adaptive broadcast in Internet of things (IoT). Broadcast is common in IoT due to the ubiquitous tree topologies. Channel resource is usually underused in broadcast because there is no rate adaptation in conventional broadcast and the data rate is always set as the lowest one by default. Existing rate adaptation methods work only for unicast or multicast, relying on information interaction between senders and receivers. These methods cannot directly apply in broadcast, which is a one-way transmission without acknowledgement (ACK). It is also impractical to transplant conventional ACK into broadcast, otherwise, massive ACKs will lead to a heavy overhead. To tackle this dilemma, we propose RAB, which allows the sender to broadcast data ceaselessly while adjusting the data rate according to real-time channel states. The core contribution is the subtly designed feedbacks that can be concurrently delivered and do not affect any reception. We implement RAB on USRPs and establish a 20-node IoT testbed. Experiment results demonstrate that the throughput is largely improved. The throughput of RAB is 2.8x of the standard WiFi and 1.3x of MuDRA, the state-of-the-art multicast rate adaptation method.

Keywords: Internet of things · Rate adaptation · Acknowledgement · Wireless broadcast

1 Introduction

Wireless broadcast is an efficient solution to deliver data that one sender (server) transmits data to all neighbors (clients) simultaneously. Its value lies in plenty of Internet of things (IoT) applications. For example, data dissemination in Internet of vehicles [19] and data sharing in collaborative robots [7].

However, the throughput of wireless broadcast is extremely low. Using 802.11g WiFi as an example, the data rate of broadcast is always set at the lowest 6 Mbps. Field tests [10] reveal that even if three clients connecting to

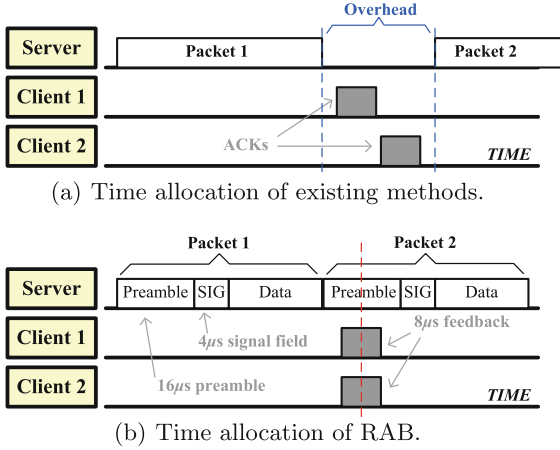


Fig. 1. Compared with existing rate adaptation methods, RAB removes the overhead.

one AP try to watch the same video, the performance is abysmally poor. The performance of wireless broadcast leaves much to be desired.

Rate adaptation is the fundamental technique to improve the throughput in dynamic wireless channel. Existing studies usually focus on unicast and multicast. In unicast, diverse rate adaptation methods have been investigated using frequency [9], constellation [11], collision [5], or SNR [18] analysis. All these methods depend on the information exchange between sender and receiver. In multicast, recent studies allow only partial clients to reply in order to balance accuracy and overhead. For example, in REMP [8], only NACKs are sent. In MuDRA [3], the server collects ACKs from representative clients by a lightweight protocol. Nevertheless, these methods are impractical in broadcast because: (i) there is no ACK in broadcast; (ii) the number of clients may be large and their channel states are different. If the conventional ACK mechanism is adopted, heavy overhead will be introduced due to a large amount of clients, largely reducing the throughput.

We observe that the preamble in WiFi has opportunity to improve this problem. The essence of preamble is a training sequence at the packet header. Even partial samples in the preamble cannot be decoded, the packet still can be successfully received. Our main idea is that the server keeps broadcasting packets while all clients concurrently transmit their feedbacks for rate adaptation during the preamble time as shown in Fig. 1.

There are two major challenges to realize this idea. First, since all feedbacks and the preamble are parallel, they are collided together. Processing this overlapped signal is difficult because the interference is from not only the server but also all clients. Second, the duration of preamble is only $16\mu\text{s}$ in WiFi, equal to four OFDM symbol length. When numerous clients exist, it is not easy to design such short feedbacks containing all required and decodable information.

Table 1. IEEE 802.11a/g data rate information.

Modulation	Coding	Data rate	Effective SNR
BPSK	1/2	6 Mbps	<6.6 dB
BPSK	3/4	9 Mbps	6.6–8.7 dB
QPSK	1/2	12 Mbps	8.7–9.6 dB
QPSK	3/4	18 Mbps	9.6–17.3 dB
16QAM	1/2	24 Mbps	17.3–18.4 dB
16QAM	3/4	36 Mbps	18.4–26.0 dB
64QAM	2/3	48 Mbps	26.0–28.1 dB
64QAM	3/4	54 Mbps	>28.1 dB

To tackle these challenges, we propose the *rate adaptive broadcast* (RAB). Fully exploiting the WiFi subcarrier, every client just transmits a two-symbol-length feedback, while the server extracts the needed information from the collisions of all clients. The other advantages of RAB include: the subtly designed feedback has the same structure of standard OFDM symbols. So it is easy to be implemented and is compatible to commercial WiFi devices; RAB is also a general solution, which can be extended to other wireless protocols.

The main contributions of this paper are as follows:

- We design RAB that enables rate adaptive broadcast in IoT. The core design of RAB leverages the preamble to realize the concurrent transmission and leverages the orthogonal subcarriers to develop the short feedbacks.
- We implement RAB on USRPs, and establish a 20-node testbed. Experiment results demonstrate that RAB achieves the mean throughput gain of 2.8x and 1.3x compared with the standard WiFi and the state-of-the-art MuDRA, respectively.

2 Problem Statement

Before introducing RAB, we review the background and understand the WiFi preamble. Then, we state our problem and summarize the design challenges.

2.1 Background

Rate adaptation is a fundamental primitive in wireless networks. Since the channel state varies unpredictably, a sender has to measure the dynamic channel and selects the appropriate data rate to maximize the throughput. Taking IEEE 802.11a/g WiFi as an example, eight different data rates are available including 6, 9, 12, 18, 24, 36, 48, 54 Mbps. We list the data rate information in Table 1 and with the effective SNR measured by our experiment (details in Sect. 4).

Broadcast, multicast, and unicast are three general communication modes in WiFi. If clients have interests in the same content, broadcast is the most

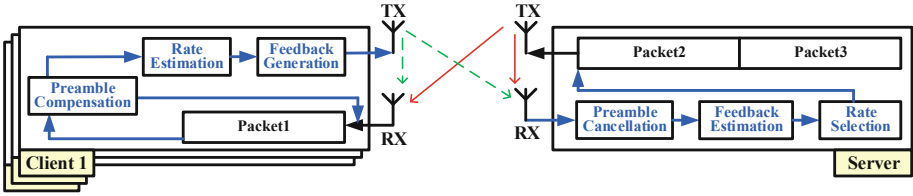


Fig. 2. RAB architecture.

efficient mode that utilizes one channel to transmit data to all clients. In IEEE 802.11a/g, if the broadcast mode is chosen, the data rate is fixed at the lowest 6 Mbps. Due to no ACK in broadcast, the server cannot acquire different channel states from all clients, so the lowest rate is the conservative setting.

2.2 Understanding of WiFi Preamble

Preamble is a pre-defined training signal at the packet header. Nearly all wireless protocols require preambles, because packet detection and time synchronization between sender and receiver totally rely on it. WiFi's preamble is $16\ \mu\text{s}$ length consisting of an $8\ \mu\text{s}$ *short training field* (STF) and an $8\ \mu\text{s}$ *long training field* (LTF). STF is the 10 repetitions of a given 16-sample sequence and LTF is the 2.5 repetitions of a 64-sample sequence, where the repetition pattern is a 32-sample *cyclic prefix* (CP) and then two 64-sample signals. When a client receives a preamble, it operates the correlation by known STF and LTF sequences. Through the 10 correlation peaks in STF and 2 correlation peaks in LTF, the client can detect the start-of-packet and complete the time synchronization.

Following the preamble is the *signal field* (SIG), which contains the information of data rate and packet length. Specially, the *reserved* bit is intended for future use.

2.3 Problem, Challenges, and Observations

In IoT, the lowest data-rate broadcast obviously lowers the quality of experience. Motivated by fully exploiting the channel resource and improving the throughput, we propose to study the rate adaptation problem in WiFi broadcast.

Rate adaptation and broadcast never work collaboratively in standard WiFi. On one hand, an accurate adaptation depends on the two-way interaction to measure all channel states. However, the broadcast is a one-to-many and one-way transmission mode without ACKs. On the other hand, existing reactive rate adaptation methods, used in unicast and multicast, cannot directly apply in broadcast. Since there may be numerous clients in broadcast case, one-by-one ACKs result in a heavy overhead.

Inspired by full duplex [1] and concurrent transmission [12], we conceive a concurrent-feedback design to tackle the above dilemma. In addition, we observe two opportunities from the packet structure, which are able to facilitate the

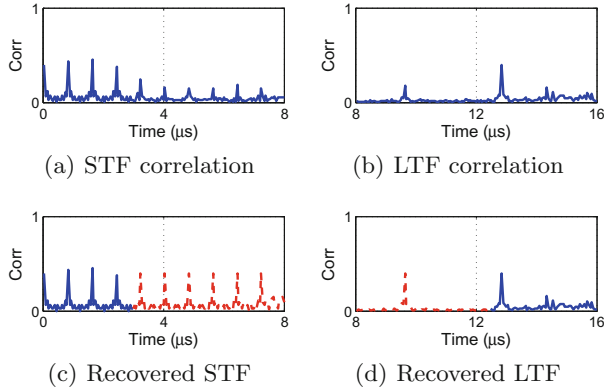


Fig. 3. The correlation and recovered results on the collided signal.

design. First, the essence of the preamble is a training sequence. Partial missing preamble will not lead to any data loss. Moreover, the missing part is potential to be compensated by the known sequence and channel coherence. Second, the reserved bit in signal field can be used to transmit 1-bit information.

3 Design of RAB

Based on the observations, we design the *rate adaptive broadcast* (RAB) to bridge rate adaptation and broadcast in IoT. The block diagram of RAB architecture is shown in Fig. 2, including the designs of client and server. In RAB, every device equips one TX and one RX antennas, which is a usual configuration in commercial WiFi devices.

- At the client side, three modules are added. After receiving a packet, the *preamble compensation* module is ready to compensate the collided preamble of next packet; while the *rate estimation* module estimates the supported data rate by analyzing the received packet. Then, the *feedback generation* module generates the short RAB feedback implying the estimated rate and transmit it to the server during the preamble time.
- At the server side, three modules are added. The server receives an overlapped signal containing the clients' feedbacks and its own preamble. The function of *preamble cancellation* filters the feedbacks out from the preamble. However, the multiple feedbacks are still overlapped. So the *feedback estimation* module recognizes every client's information from the overlapped feedbacks. According to the recognized information, the *rate selection* determines the optimal data rate.

The greatest strength of RAB is to enable the rate adaptation in broadcast without extra time overhead. The server fully exploits the channel in time domain by broadcasting data ceaselessly. On the contrary, all clients receive the data in

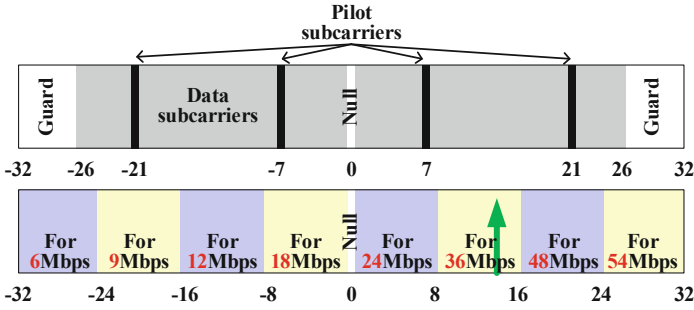


Fig. 4. Subcarrier allocation for OFDM symbol (up) and RAB symbol (down).

most time. They only concurrently transmit their $8\ \mu\text{s}$ feedbacks within the $16\ \mu\text{s}$ preamble time (align center).

It is not easy to realize RAB. The first challenge is the collision resolution. The severe collision caused by the parallel server’s preamble and clients’ feedbacks. Another challenge is how to design the short feedback so that the server can recognize the information from the parallel feedbacks. Next, we describe the design of every module in details.

3.1 Client: Preamble Compensation Module

Every client is only interested in the preamble, but the preamble is hard to be extracted from the collided signal. Fortunately, unlike the payload, the preamble is a training sequence attaching no essential data. Hence, a client just needs to compensate the preamble.

First, for start-of-packet detection and time synchronization tasks, the *preamble compensation* module directly operates the correlation on the collided signal by 16-sample STF sequence and 64-sample LTF sequence. The correlation results are shown in Fig. 3(a) and (b). Although several correlation peaks are distorted, their number and locations are clearly recognized. Hence, the detection and synchronization tasks can be accomplished as usual.

Then, for AGC, this module leverages the feature that the received preamble is partially overlapped. Every feedback signal is $8\ \mu\text{s}$ -duration overlapped at the center of the $16\ \mu\text{s}$ preamble. So the first 80 samples in STF and last 80 samples in LTF are nearly non-collided. To be conservation, we use the first correlation peak in STF and the last correlation peak in LTF to recover the other peaks in Fig. 3(c) and (d). Therefore, AGC task is accomplished.

3.2 Client: Rate Estimation Module

The *rate estimation* module aims to assess the maximal supported data rate. In literature, plenty of metrics have been adopted to assess the data rate such as constellation [11], collision [5], and SNR [18]. This module is open to any existing

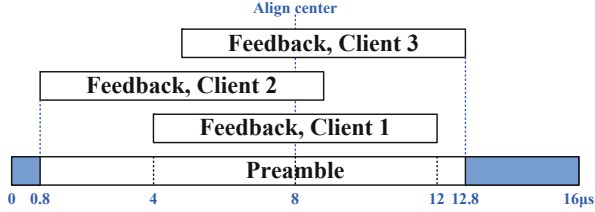


Fig. 5. The synchronization goal is the ‘align center’ between preamble and RAB feedback. However, it can be tolerant as long as the feedback falls in 0.8–12.8 μs range.

method as long as the data rate could be accurately and quickly estimated. In current RAB, we adopt the classic metric *effective SNR* [4]. Yet SNR is coarse and insufficient, the effective SNR is more accurate to adapt the rate in dynamic channel.

When a client receives the packet, the maximal supported rate is obtained by the following steps: (i) calculate CSI by the received packet (ii) the MMSE expression is used to compute subcarrier SNRs from the CSI; (iii) the effective SNR is computed from the subcarrier SNRs; and (iv) assess the maximal supported data rate through the effective SNR.

Since the effective SNR is a mature technique, we just briefly introduce how to compute the effective SNR. First,

$$\mathcal{B}_{eff,k} = \frac{1}{52} \sum_{s=-26}^{26} f_{\mathcal{S} \rightarrow \mathcal{B},k}(\mathcal{S}_s), \quad (1)$$

where $\mathcal{B}_{eff,k}$ is the effective BER at the pre-set data rate k , $k \in \{6, 9, \dots, 54\}$, s is the indicator of subcarriers belonging to $[-26, 26]$ and $s \neq 0$, $f_{\mathcal{S} \rightarrow \mathcal{B},k}(\cdot)$ is the mapping function from SNR to BER, and \mathcal{S}_s is the SNR of the s -th subcarrier. According to different k , $\mathcal{B}_{eff,k}$ needs to be computed respectively because of different mapping function. For example, if $k = 6$ Mbps, $f_{\mathcal{S} \rightarrow \mathcal{B},k}(\mathcal{S}_s) = Q(\sqrt{2\mathcal{S}_s})$, where Q is the standard normal CDF; if $k = 48$ Mbps, $f_{\mathcal{S} \rightarrow \mathcal{B},k}(\mathcal{S}_s) = \frac{7}{12}Q(\sqrt{\mathcal{S}_s/21})$; the mapping functions for the other data rates can be found in [4]. Then,

$$\mathcal{S}_{eff,k} = f_{\mathcal{B} \rightarrow \mathcal{S},k}(\mathcal{B}_{eff,k}), \quad (2)$$

where $\mathcal{S}_{eff,k}$ is the effective SNR at data rate k and $f_{\mathcal{B} \rightarrow \mathcal{S},k}(\cdot)$ is the inverse function of $f_{\mathcal{S} \rightarrow \mathcal{B},k}(\cdot)$.

3.3 Client: Feedback Generation Module

After the data rate is estimated, the client needs to generate the feedback containing this data rate information. Since a server cannot decomposed the collision of conventional ACKs, it is necessary to design a novel pattern of feedback, refer to *RAB feedback*. In RAB, the feedbacks are overlapped in time domain, we conceive to distinguish them from the frequency domain.

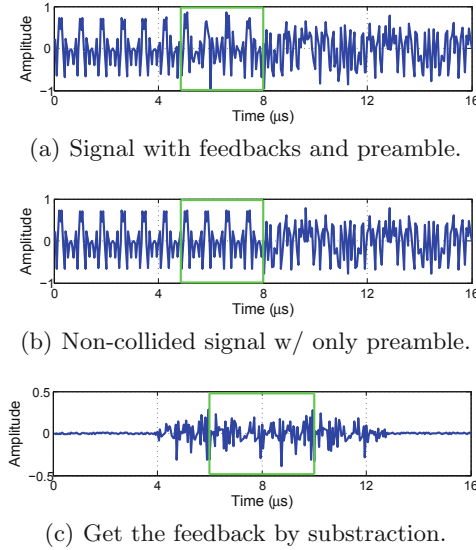


Fig. 6. The process of preamble cancellation.

Let us first understand the OFDM symbol, which is the least transmission unit. Each OFDM symbol consists of 64 orthogonal subcarriers in frequency domain and 80 samples in time domain. The allocation of these subcarriers is illustrated in Fig. 4(up).

Based on the OFDM symbol, we design the *RAB symbol*, a customized symbol for RAB feedback. Each RAB symbol also consists of 64 subcarriers. The allocation of these subcarriers is illustrated in Fig. 4(down), where 64 subcarriers are evenly divided into 8 groups mapping to 6, 9, 12, 18, 24, 36, 48, 54 Mbps, respectively. According to its maximal supported data rate, a client randomly selects one subcarrier within the corresponding group to transmit a peak and all the other subcarriers are set null. Then, operating IFFT on the RAB symbol, the 64 corresponding samples are obtained in time domain. Repeating 2.5 times of these samples, we have the 8 μs RAB feedback (160 samples).

The advantages of RAB feedback include: (i) The RAB symbol is compatible to commercial WiFi devices, because the framework of RAB symbol is the same as OFDM symbol. (ii) All 64 subcarriers are fully exploited to maximize the number of different feedbacks. The guard and pilot subcarriers are unnecessary because the feedback is short and simple. (iii) The length of feedback is minimized in order to reduce the interference on preamble.

Besides generating the RAB symbol, another job of the *feedback generation* module is to synchronize the feedback transmission. The goal is to align center between the preamble and the feedbacks as shown in Fig. 5, i.e., transmitting every feedback at the 4–12 μs position of the preamble. So that the first 4 μs of STF and the last 4 μs LTF can be non-collided for training tasks. This goal is

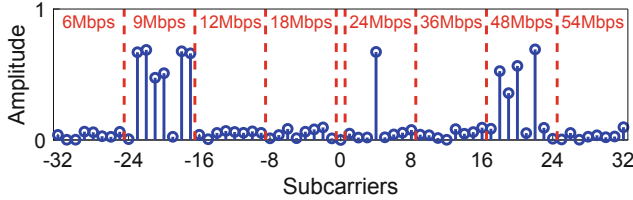


Fig. 7. Transform the feedback signal from time to frequency domain.

approached by two steps. First, since the finish time of current packet reception is known as t and the server broadcasts its packets ceaselessly, the feedback can be transmitted at $t + 4\mu\text{s}$. Second, the client could detect the time offset between the preamble and its own feedback.

Even utilizing these two steps, the synchronization may be not perfect because of hardware limitation. Fortunately, some offset can be tolerant in our design. We find that the first repetition of STF sequence finishes at $0.8\mu\text{s}$ and the last repetition of LTF sequence starts at $12.8\mu\text{s}$. Thus, as long as the RAB feedback is transmitted within the range of $0.8\text{--}12.8\mu\text{s}$, the aforementioned two repetitions are still non-collided, which is sufficient to accomplish the training tasks. Compared with the ideal $4\text{--}12\mu\text{s}$ position, the tolerated offset is up to $4\mu\text{s}$. Such a offset is easily achievable by existing technique [15], which can realize a $<0.5\mu\text{s}$ synchronization for concurrent transmissions.

3.4 Server: Preamble Cancellation Module

As shown in Fig. 6, the server receives a collided signal, which is dominated by its own preamble. The *preamble cancellation* module aims to cancel this preamble and filter the feedbacks out. To operate the cancellation, we introduce the non-collided preamble in the last packet. Due to the channel coherence and two consecutive packets, the last preamble is potential to be the baseline to cancel the preamble in current collision. Nevertheless, since the phase offset may exist in different complex signals, we cannot subtract the non-collided signal directly from the collided signal.

The *preamble cancellation* module operates the cancellation in frequency domain. A 64-sample time window is set to extract the time-domain signal in both the collided and non-collided signals, where the 64-sample is the least window to do FFT on 64 subcarriers. To guarantee that all feedbacks have at least 64-sample overlapped together, the minimal duration of a feedback is $8\mu\text{s}$. And the location of time window is from 4.8 to $8\mu\text{s}$. Hence, the information of all clients are included in this time window.

Operating FFT on both extracted signals and doing the cancellation by magnitude subtraction, the preamble signal is cancelled and we have all clients' feedbacks in frequency domain as shown in Fig. 7.

According to the subcarrier allocation in RAB symbol, the distribution of peaks can be counted in every data rate. Since the noise can be easily measured,



Fig. 8. RAB testbed.

a peak is determined when its amplitude is larger than the average noise. We note a subcarrier with peak as ‘1’ and a subcarrier without peak as ‘0’.

3.5 Server: Feedback Estimation Module

A client randomly selects one subcarrier in the group of its data rate. Hence, it is possible that multiple clients select the same subcarrier, especially in dense case. Considering this case, the goal of the *feedback estimation* module is to compute the number of clients in every group of data rate. Various existing methods serve for cardinality estimation based on responses of ‘0’ and ‘1’. For example, UPE [6] and ART [12]. This module is open to diverse such methods as long as the number of clients can be accurately estimated.

In current RAB version, we adopt the classic *unified probabilistic estimation* (UPE) [6] to realize this feedback estimation module. Denote n_0 as the number of ‘0’ s, m as the number of subcarriers in a group, and N as the number of clients in this group. UPE estimates N on account of n_0 and m . The UPE estimator is

$$\tilde{N} = -m \times \ln\left(\frac{n_0}{m}\right), \quad (3)$$

where \tilde{N} is the estimate of N . According to Eq. (3) and the setting of $m = 64/8 = 8$, in Fig. 7, there are 6 peaks can be found in the group of 9 Mbps, so $n_0 = m - 6 = 2$. Then, UPE estimates the number of clients $\tilde{N} = -8 \times \ln(2/8) \approx 11$.

3.6 Server: Rate Selection Module

Using the result of feedback estimation, the *rate selection* module determines the optimal data rate according to the applications. Various policies can be defined such as ‘maximize the throughput’ or ‘maximize the total number of clients’.

To see how rate selection module works, we use a simple example. If the policy is ‘maximize the total number of clients’, 9 Mbps is the optimal data rate, in which all clients could decode the broadcasting data from server. If the policy is ‘maximize the total throughput’, 48 Mbps is the optimal rate. Although only 5 clients achieve the successful reception, the total throughput is $48 \times 5 = 240$ Mbps, which is larger than selecting 9 or 24 Mbps.

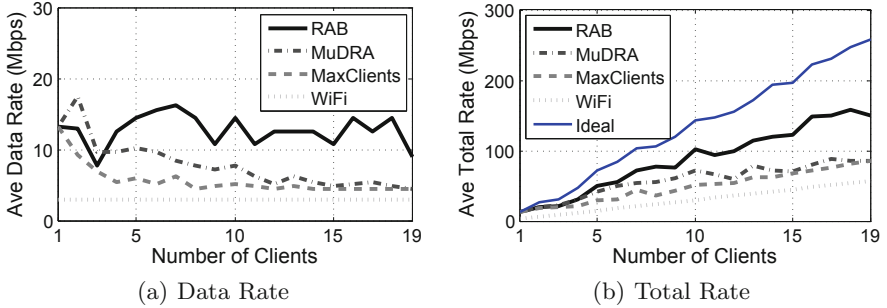


Fig. 9. Rate adaptation results.

4 Performance Evaluation

We implement RAB on USRP and build a 20-node testbed to evaluate its performance.

4.1 Implementation

Platform: We build the prototype of RAB using the GNU Radio toolkit and USRP N210. All the USRP nodes are equipped with SBX daughter boards and two VERT2450 antennas, so that they can simultaneously transmit and receive on 2.4 GHz. We adopt existing *IEEE 802.11 a/g/p transceiver for GNU Radio* [2] as the basic WiFi physical layer (PHY).

RAB PHY: The design of RAB PHY is shown in Fig. 2. We implement it according to the design. To trigger the rate adaptation, the server sets the *reserved bit* as ‘1’. The server keeps broadcasting packets while it logs the non-collided preamble for the use of preamble cancellation. Besides, every client estimates the noise level from the received packet and keeps updating the average SNR for each subcarrier in order to calculate the effective SNR.

MAC: Since RAB works at the broadcast mode, the carrier sensing in MAC layer is disabled in our prototype.

SNR: The relationship between the maximal supported rate and the effective SNR is measured using our USRP nodes. Then, every client can empirically determine its maximal supported rate when the effective SNR is obtained.

Testbed: As shown in Fig. 8, our testbed includes 20 USRP nodes as an IoT covering an office, whose area is 32 m². One server and 19 clients build a single-hop topology for data broadcasting.

Configuration: All USRP nodes operate at 2.484 GHz. To avoid the odd decimation, we use the bandwidth of 10 M instead of 20 M. Hence, the data rates are halved, *e.g.*, the lowest data rate is 3 Mbps in our evaluation. The size of every packet is 1500 Byte.

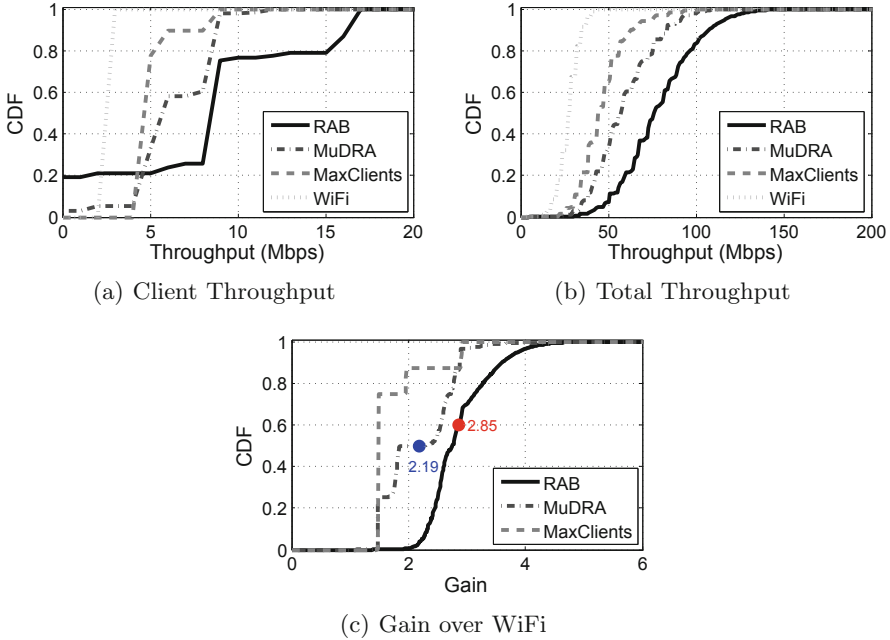


Fig. 10. Comparison on throughput.

Compared Schemes: We compare RAB with

- MuDRA [3]: is the state-of-the-art rate adaptation design for multicast, which adapts the data rate by lightweight feedbacks from partial clients. MuDRA aims to ensure $>85\%$ packet reception ratio (PRR) for $>95\%$ clients. The feedback interval is set to be 500 ms as used in [3], and no more than 4 clients will send their feedbacks every 500 ms.
- MaxClient: is a straightforward method that selects the data rate for maximizing the number of clients.
- WiFi: is the standard IEEE 802.11a/g protocol. The server broadcasts with the lowest data rate and the clients do not transmit ACK.

4.2 Evaluation Result

Micro Benchmark Rate Adaptation. The server broadcasts 1500Byte packets to a client with a 50 ms interval using RAB, MuDRA, MaxClient, and WiFi. The aim of such real-time experiments is to validate the rate adaptation in RAB. In the experiments, the USRP nodes are kept stationary, leading to relatively stable effective SNR.

Figure 9 shows the average data rate when the number of clients increases from 1 to 19. We observe that RAB always selects the highest data rate, because

it does not sacrifice the overall throughput for poor clients. To be specific, when the number of clients increases, MaxClients tends to be stuck in the data rate of 4.5 Mbps, which is the lowest data rate for all 19 clients. Meanwhile, since MuDRA is allowed to ignore the worst client, it is slightly better than MaxClients, but still fall behind RAB. In Fig. 10, we adopt the sum of each client’s data rates, i.e., the total rate, to indicate the throughput. The high total rate implies that RAB scales well with increasing clients. These results verify the promising advantage of RAB in broadcast.

System-Level Gain. Based on the experiment results, we collect the maximal supported data rates for all the 19 clients for system-level simulation. For comparison, we assume MaxClients utilize the feedbacks without extra overhead, while the standard WiFi directly broadcasts packets using the lowest data rate. Moreover, as [3], each feedback in MuDRA is 1 ms, and we let all the feedback nodes transmit successively.

We iterate all the possible combinations of the 19 clients, and calculate the cumulative distribution of the throughput shown in Fig. 10. From Fig. 10(a), we find that in WiFi, all clients have the same throughput because they all receive packets from the server at the lowest data rate. Besides, both RAB and MaxClients have partial zero throughput. The reason is that the server “abandons” some clients with poor link quality. However, those zero-throughput clients (less than 20%) does not impact the network-wide throughput. As shown in Fig. 10(b), the overall throughput gain of RAB is tremendous, *i.e.*, the median throughput increments compared with MuDRA and MaxClients are as high as 31.03% and 72.73%, respectively.

Specifically, we evaluate the total throughput gains of RAB, MaxClients, and MuDRA, compared with the standard WiFi broadcast in Fig. 10(c). The median gain of RAB and MuDRA are 2.76x and 1.87x respectively. Moreover, RAB can achieve an average throughput gain of 2.85x compared with WiFi, which is also 30.51% higher than that of MuDRA.

5 Related Work

We classify existing rate adaptation techniques into two categories.

Rate Adaptation in Unicast. Adapting the suitable data rate to the dynamic channel is valuable for wireless communication, where a too high data rate leads to decoding error and a too low rate wastes the channel resources. In unicast, the data rate is adjusted by various metrics. RRAA [16] measures the packet loss rate. SoftRate [14] utilizes the SoftPHY hints to obtain the per packet BER. FARA is a frequency awareness rate adaptation [9]. AccuRate [11] investigates the constellation state. CARA analyzes the collisions [5]. These methods perform well in unicast. However, collecting the metrics from all clients is not practical in broadcast due to heavy overhead.

Rate Adaptation in Multicast. To improve the throughput in multicast, extensive low-overhead rate adaptation methods are developed. REMP [8]

requires the non-reception clients to send NACKs. ARSM [13] reduces the overhead by selecting partial clients to send feedbacks, typically the clients with poor channel quality. Peercast [17] improves the link layer multicast through collaborative relays and batch ACKs. MuDRA [3] dynamically adjusts the data rate relying on collecting representative feedbacks via a light-weight protocol. However, overhead of above methods cannot be completely removed.

6 Conclusion

This paper presents the novel RAB to enable the rate adaptation in IoT broadcast. Leveraging the preamble and orthogonal subcarriers, RAB collects parallel feedbacks from all clients during the preamble time, which has no affect on packet reception and requires no extra overhead. Through implementation and testbed based performance evaluation, we show that RAB embraces the parallel acknowledgements, separates feedbacks from preamble signals, and increases the throughput in broadcast.

The future work includes the RAB transplant from WiFi to other IoT wireless protocols such as ZigBee and LoRa, the RAB extension on subcarrier allocation to obtain more accurate estimation, and the RAB generalization from broadcast to multicast.

Acknowledgment. This work was supported in part by the National Key R&D Program of China 2018YFB1004703, National Natural Science Foundation of China grant 61972253, 61672349, 61672353.

References

1. Bharadia, D., McMilin, E., Katti, S.: Full duplex radios. In: ACM SIGCOMM (2013)
2. Bloessl, B., Segata, M., Sommer, C., Dressler, F.: An IEEE 802.11 a/g/p OFDM receiver for GNU radio. In: ACM SIGCOMM Workshop on Software Radio Implementation Forum (2013)
3. Gupta, V., Gutterman, C., Bejerano, Y., Zussman, G.: Experimental evaluation of large scale WiFi multicast rate control. In: IEEE INFOCOM (2016)
4. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Predictable 802.11 packet delivery from wireless channel measurements. In: ACM SIGCOMM (2011)
5. Kim, J., Kim, S., Choi, S., Qiao, D.: CARA: collision-aware rate adaptation for IEEE 802.11 WLANs. In: IEEE INFOCOM (2006)
6. Kodialam, M., Nandagopal, T.: Fast and reliable estimation schemes in RFID systems. In: ACM MobiCom (2006)
7. Kong, L., et al.: Adasharing: adaptive data sharing in collaborative robots. *IEEE Trans. Ind. Electron.* **64**(12), 9569–9579 (2017)
8. Lim, W.-S., Kim, D.-W., Suh, Y.-J.: Design of efficient multicast protocol for IEEE 802.11n WLANs and cross-layer optimization for scalable video streaming. *IEEE Trans. Mob. Comput. (TMC)* **11**(5), 780–792 (2012)
9. Rahul, H., Edalat, F., Katabi, D., Sodin, C.: Frequency-aware rate adaptation and MAC protocols. In: ACM MobiCom (2009)

10. Sen, S., Madabhushi, N.K., Banerjee, S.: Scalable WiFi media delivery through adaptive broadcasts. In: NSDI (2010)
11. Sen, S., Santhapuri, N., Choudhury, R.R., Nelakuditi, S.: AccuRate: constellation based rate estimation in wireless networks. In: NSDI (2010)
12. Shahzad, M., Liu, A.X.: Every bit counts: fast and scalable RFID estimation. In: ACM MobiCom (2012)
13. Villalón, J., Cuenca, P., Orozco-Barbosa, L., Seok, Y., Turletti, T.: Cross-layer architecture for adaptive video multicast streaming over multirate wireless lans. *IEEE J. Sel. Areas Commun. (JSAC)* **25**(4), 699–711 (2007)
14. Vutukuru, M., Balakrishnan, H., Jamieson, K.: Cross-layer wireless bit rate adaptation. In: ACM SIGCOMM (2009)
15. Wang, Y., He, Y., Mao, X., Liu, Y., Li, X.-Y.: Exploiting constructive interference for scalable flooding in wireless networks. *IEEE/ACM Trans. Netw. (ToN)* **21**(6), 1880–1889 (2013)
16. Wong, S.H., Yang, H., Lu, S., Bharghavan, V.: Robust rate adaptation for 802.11 wireless networks. In: ACM MobiCom (2006)
17. Xiong, J., Choudhury, R.R.: PeerCast: improving link layer multicast through cooperative relaying. In: IEEE INFOCOM (2011)
18. Zhang, J., Tan, K., Zhao, J., Wu, H., Zhang, Y.: A practical SNR-guided rate adaptation. In: IEEE INFOCOM (2008)
19. Zhou, Z., Gao, C., Xu, C., Zhang, Y., Mumtaz, S., Rodriguez, J.: Social big-data-based content dissemination in internet of vehicles. *IEEE Trans. Ind. Inform.* **14**(2), 768–777 (2017)



An IOT Data Collection Mechanism Based on Cloud-Edge Coordinated Deep Learning

Zi-hao Wang and Jing Wang^(✉)

Beijing Key Laboratory on Integration and Analysis of Large-Scale Stream Data,
Data Engineering Institute, North China University of Technology,
Beijing, China
13681272894@163.com, wangjing@ict.ac.cn

Abstract. The large-scale of data collection from IoT devices to central cloud brings several challenges that need to be overcome, especially for needs of real-time collection and bandwidth restrictions. In order to address this issue, we proposed a data collection method that combine cloud with edge node by using deep learning technology to provide the data collection service. The cloud is responsible for storing the large amount of historical sensor data, training the deep learning model, and deploying the model to the edge side. The edge node will receive the model of data prediction and then determines whether the real data will be uploaded to the cloud to optimize the model. Experiments show that the method we proposed can not only increase the speed of data collection, but also reduces the network traffic and eliminates bandwidth load effectively.

Keywords: Data collection · Cloud-edge coordination · Deep learning

1 Introduction

With the development of Internet of Things (IoT) technology, IoT system will face a lot of data due to the numerous smart terminal devices. According to research firm Gartner, as many as 20 billion connected devices will generate billions of bytes of data per user by 2020. These devices are not just smartphones or laptops, but also connected cars, vending machines, smart wearables, surgical medical robots, and more. The vast amount of data generated by countless types of such devices needs to be pushed to a centralized cloud for retention (data management), analysis and decision making. The analyzed data results are then passed back to the device if needed. This round-trip of data consumes a large amount of network infrastructure and cloud infrastructure resources, further increasing latency and bandwidth load issues, thereby affecting critical mission of IoT system [1]. In this case, edge computing technology is rapidly integrated into large-scale IoT monitoring systems [2] and is supported by edge computing devices such as smart gateways, lightweight servers, and small base stations. In the IoT system, the edge computing layer is closer to the IoT terminal devices, therefore, it can not only provide local data collection and improve the real-time performance of the service, but also reduce the backhaul delay between the terminal devices and the cloud platform. However, edge nodes have very limited computing, storage, and network resources, which make it difficult to process large-scale data.

In this paper, a method of data collection using deep learning technology combining cloud and edge nodes is proposed. Using advantage of computing power of the cloud and real-time superiority of the edge node, the neural network model can be trained in cloud and deployed to the edge node and predict sensor data at the edge node. When the prediction accuracy is less than the threshold, the prediction result will be saved in the edge node as a match object for the next business; otherwise the real data will be uploaded to the cloud and will be used to optimize the training model. The experiment results show that this method can not only increase the speed of computing services and satisfy the real-time requirements of data collection, but also reduces network traffic and eliminates bandwidth load effectively.

The rest part of this paper is organized as follows: Sect. 2 introduces related research on the edge computing in the field of deep learning. Section 3 mainly introduces the principle of the method proposed in this paper and the data processing flow. Section 4 introduces the realization of the test system. Section 5 presents the experimental design and verification of results. Section 6 makes a conclusion and outlook for future work.

2 Related Work

The ambitious vision of IoT not only led to many studies in scientific and academic communities, but also attracted many industrial domains. Considering the overall methodological perspective of IoT, there are still problems for the IoT data collection. Firstly, the delay is too long caused by the distance between the IoT device and the central cloud. Secondly, pushing a large amount of sensor data packets to the cloud would increase the bandwidth load of the network. Aiming at IoT data collection and load balancing, researchers have developed various data collection models and prediction mechanisms for IoT applications.

The first method is to push the sensor data into the central cloud for processing directly (Sensor-Cloud). Madria et al. [3] have proposed a sensor cloud architecture that connects different wireless sensor networks in vast geographic areas. This architecture can be used by multiple users “on demand” at the same time. Virtual sensors will help create a multi-user environment based on resource-constrained physical wireless sensors and help support multiple applications on demand. However, Madria does not consider the high network load and delay problem when the sensor transmits large amount of data. Truong et al. [4] present the design of an Internet of Things (IoT) system consisting of a device capable of sending real-time environmental data to cloud storage and a machine learning algorithm to predict environmental conditions for fungal detection and prevention. They built a data collection cluster for sensor data access and processed raw environmental data and predict short-term temperatures in the cloud using SVMr (Support Vector Machine regression) algorithm, increase throughput while helping to predict the presence of harmful fungi and prevent disease collection. However, they do not consider the coordination between cloud and edge node, the number of packet collections cannot be further reduced.

The second method adds an edge node as the middle layer based on the first method (Sensor-Edge-Cloud). Peralta et al. [5] extend MQTT, the de facto IoT collection protocol, using an edge computing approach that introduces a low complexity computational layer between the Cloud and IoT nodes. In this approach, the MQTT broker, which is in charge of relaying data from publishers to subscribers, is placed at the edge layer. With this architecture, the collections required from IoT devices may be reduced, since the publishers would only need to update the predicted data in case of mismatching. However, their architecture is limited by the computational power of edge nodes. Oma et al. [6] propose a linear IoT model to deploy processes and data to devices, edge nodes, and servers in IoT. Edge nodes not only receive sensor data, but also do some computation and storage on a collection of data sent by sensor nodes so that the total electric energy consumption of nodes can be reduced. The experiments show the total electric energy of the IoT model is smaller than the traditional cloud model.

In the field of deep learning, many researchers have proposed different prediction algorithms with the above purpose. Hinton [7] and others use the Deep Belief Network (DBN) structure to form a DBN. Each layer of Restricted Boltzmann machine (RBM) structure was used for unsupervised learning training and used in MNIST handwritten digit recognition tasks, which achieving the best score of 1.2% error. Xiaoyun [8] et al. applied LSTM neural network to short-term wind power data prediction based on Principal Component Analysis, and the prediction error was lower than that of SVM model. The smart home machine learning system proposed by Wei [9] used the single hidden layer BP neural network (BPNN) prediction model as the core to receive environmental data and predict the data. The prediction model uses Umass Trace Repository and the prediction accuracy on the platform dataset is 85%.

Above methods provide some references for the conception of our method to solve the problem of current data collection about IoT. This paper proposes a deep learning method combining cloud with edge node. It can train the model using Bi-directional LSTM (Bi-LSTM) and send the model to edge node, the future sensor data will be predicted by using model in the edge node. The experiment results show that the method can not only increase the speed of data processing and satisfy the real-time requirements of intelligent services, but also reduces network traffic and eliminates bandwidth restrictions effectively.

3 System Architecture

The system is divided into three parts: the IoT device layer, the Edge layer, and the Cloud layer, as Fig. 1 shows. The devices send sensing data to the edge node, and real-time calculation is performed on the edge node according to the requirements. Then, the calculation results are sent to the cloud to be stored. The cloud is also responsible for computational tasks with large computational complexity and low real-time requirements, such as neural network model training tasks, and training results are transmitted to the edge.

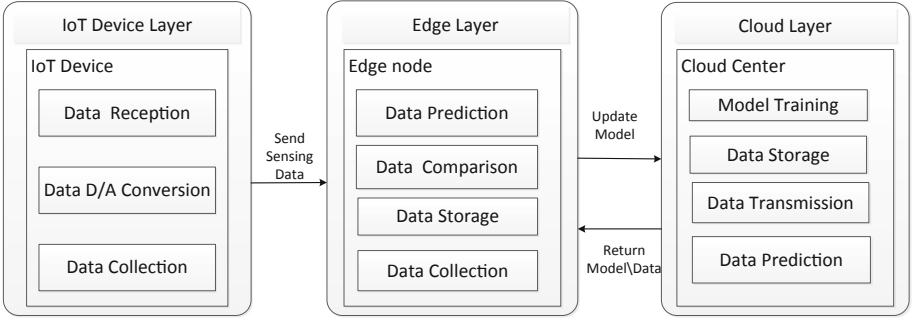


Fig. 1. Cloud-edge IoT architecture

- **IoT Device Layer:** In this layer, the sensor device nodes need to receive sensor data from current environment and push sensor data to the edge layer.
- **Edge Layer:** This layer is a real-time data storage and processing center, which is responsible for receiving real-time data transmitted from the IoT Device Layer. The Edge Layer is also responsible for calculating the accuracy value between the predicted data and the real data, and determining whether to upload real data to the cloud based on the accuracy value.
- **Cloud Layer:** This layer interacts with the edge layer and performs the historical data storage, training, prediction and deployment of deep learning models.

4 Principle of the Method

4.1 Bi-LSTM RNN

Since the IoT sensor data is mostly time series data, and the Recurrent Neural Network (RNN) can achieve high precision if the sequential machine learning task is involved. LSTM RNN is a special kind of RNN, which are characterized by the addition of valve nodes of each layer outside the RNN structure and long-term dependency problems can be avoided and long-term storage memory information can be supported by analyzing the overall logical sequence between input information.

A typical LSTM unit is shown in Fig. 2. It contains one or more memory cells c with internal states, an input gate i_t , a forgotten gate f_t , and an output gate o_t , assuming c_t is the state of memory cells at time T . Then, the calculation process of the LSTM unit at time T is as follow:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (2)$$

$$o_t = \sigma(w_o * [h_{t-1}, x_t] + b_o) \quad (3)$$

$$\tilde{c}_t = \tanh(W_c * [h_{t-1}, x_t] + b_c) \tag{4}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \tag{5}$$

$$h_t = o_t * \tanh(c_t) \tag{6}$$

In the equation: W and b are parameters to learn [9]; σ and \tanh are the sigmoid function and the hyperbolic tangent activation function, \tilde{c}_t is a new candidate value vector created by the tanh layer that will be added to the cell state, and h_t is the final output value. The LSTM model training process uses a BPTT algorithm which is similar to the classical back propagation algorithm (BP) principle [10]. The algorithm can be divided into four process steps:

- (a) Calculating the LSTM cell output value according to the forward calculation method ((1)–(6));
- (b) Calculating the error term of each LSTM cell in reverse;
- (c) Calculating the gradient of each weight according to the corresponding error term;
- (d) A gradient-based optimization algorithm is applied to update the weights.

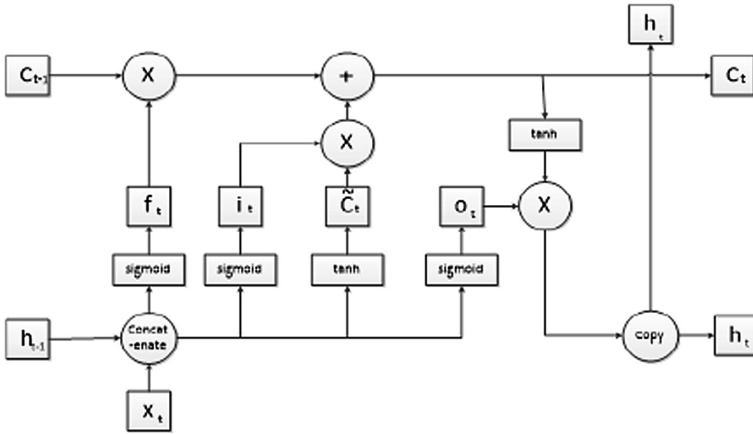


Fig. 2. LSTM hidden layer cell structure

Both RNN and LSTM can predict the output of the next moment based on the timing information of the previous moment. However, the output of the current moment is not only related to the previous state, but also may be related to the future state in some cases. In continuously changing weather data, the data at time $v1$ is not only related to historical weather data, but also related to the trend of data in a range of time in the future. The basic idea of the Bidirectional RNN [11, 12] is to divide the part responsible for the forward state and the part responsible for the reverse state in each training sequence into two RNNs, and both are connected to an output layer. This structure provides complete past and future context information for each point in the output layer input sequence.

4.2 Cloud-Edge Collaboration Algorithm

The data processing and collection flow in our proposed architecture as shown in Fig. 3. The Cloud Layer stores a large amount of historical sensor data and uses Bi-LSTM RNN neural network to train historical data, and deploys the obtained model as input data to the Edge Layer so that the edge computing layer can predict the future data. Since the time of data collection from the IoT Device Layer to the Edge is shorter than to the Cloud, the edge computing layer, which collects data once per second, first receives the raw sensor data $T(x1, x2, \dots, xn)$ at $t0$ time from the device layer and pre-processes and detects whether the local database has matching data within a period of time. If there has not matching data from database, the predicted model will execute and output the predicted data. If there has matching record in the database, the data will be compared with the predicted data which at the same time, if the difference between the predicted value and the actual value is less than the threshold, the system continues to save the data and does not modify the model which used to next prediction, otherwise the edge layer publishes the real data to the cloud every minute and the model will be trained and updated using the new data.

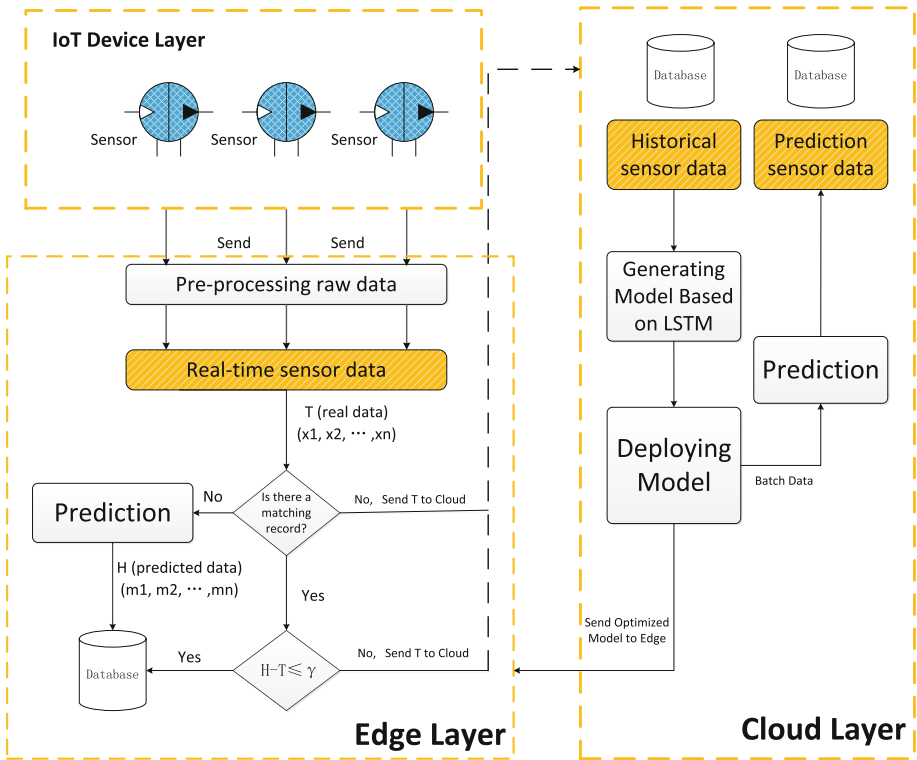


Fig. 3. Data processing and collection

Based on the above workflow, the algorithm pseudo code is described as follows:

Algorithm 1. Training and Deployment Algorithm
Input: Historical Sensing Data**Output:** Prediction Model**Start**

```

1. Scaled = fit_transform(input.values)
2. Reframed = series_to_supersived(scaled,
n_minutes, 1)
3. Train_X = X.reshape(sample, timestep, feature)
/*Standardized Data*/
4. Construct Model = LSTM(lstm_units, return_sequence
= False)
/*Design the LSTM layer */
5. Add_Dropout()
6. Add Layer = Dense(output_dim = 1, activation =
'sigmoid')
7. Compile loss, optimizer
8. Fit Model(train, epoch_num, batch_size)
/*Define fit function*/
9. Save.model(model.h5)
10. Sftp model.h5 /root/model.h5
/*Save model and deploy model to edge layer*/

```

End

Algorithm 2. Matching and Prediction Algorithm on Edge
Input: Real-time Sensing Data**Output:** Matching Data, Prediction Data**Start**

```

1. Set D0 , GPIO.setmode(GPIO.BCM)(t0)
/*Collection real sensor data of t0 through
GPIO*/
2. ADC.setup and GPIO.setup(D0, GPIO.IN)(t0)
3. analogVal = ADC.read(0)
4. Calculate Vr, Rt, Real_data(t0)
/* Convert the acquired signal to digital signal
using the ADC module*/
5. IF NOT EXISTS t0 = a ( SELECT t0_data FROM data-
base):
(1) Execute model
(2) upload real_data(0)
ELSE:  $rmse = \sqrt{(real\_data(t0) - predicted\_data(t0))^2}$ 
IF rmse < threshold:
save data
ELSE:
client.publish('v1/devices/me/attribute', pro-
cess.env.ATTRIBUTE)
Repeat step 1~5

```

End

5 Result and Analysis

5.1 Data and Environment

We have implemented a prototype system to test our method and the experiment measured by python scripts in the proposed environment. The test system includes above three layers we proposed and operation of each layer is as follows:

The IoT Device Layer, as shown in Fig. 4, which is in charge of collecting raw sensor data, is the bottom layer and is consist of digital temperature sensor node. Then, the raw sensor data sent to the Raspberry Pi [13] using the GPIO collection module and the received analog signal is converted to a digital signal using the ADC analog-to-digital conversion module (PCF8591).

The second layer is the Edge Layer which has been implemented using Raspberry Pi. The purpose of the edge computing layer is to reduce the amount of data which sent to the cloud and reduce the collection delay and bandwidth load. It calculates the accuracy difference between the predicted data obtained by RNN model comes from the cloud and standardized real-time data from the sensor. According to the error value, the edge node determines whether to upload the real data to the cloud.

The top layer is the Cloud Layer, which is implemented using ThingsBoard IoT cloud platform [14] (Fig. 5) running on a local server as a visualization system. We extent two tables in the ThingsBoard database, one is used to store historical sensor data and the other for prediction result. This layer not only to act as persistence storage of sensor data in a Postgresql database and train the deep learning model of these data, but also show the sensor values in more user-friendly graphs and tables. The ThingsBoard IoT cloud system has many of these functions built in, including an MQTT broker and an MQTT client to listen and handle MQTT message from the Edge Layer. Hence, the cloud system simply acts as a back-end storage system and sensor data visualizer, based on the sensor data sent from the Edge Layer.

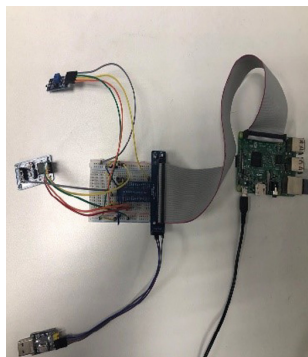


Fig. 4. Raspberry Pi and sensor

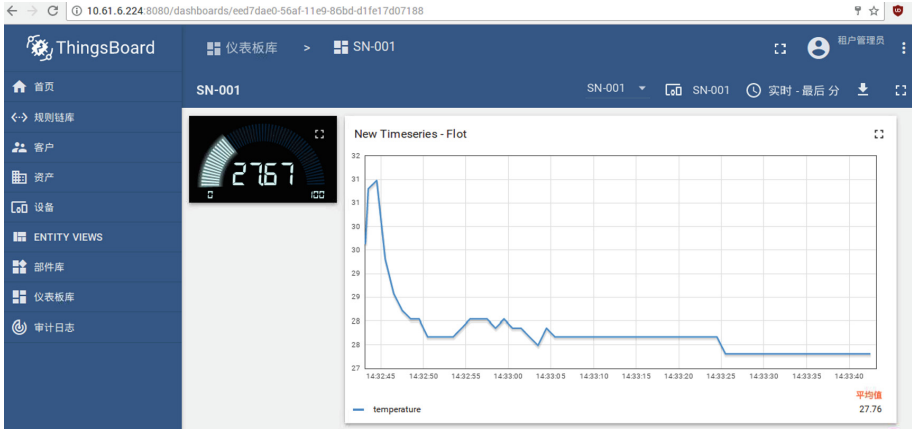


Fig. 5. Cloud dashboard

The experiment data used in this article comes from a real temperature sensor of laboratory, which including 75 h temperature sensor data and collected every one minute. First, we extract the temperature data and integrated data into CSV format. Then, the timestamp dimension is deleted, and all the raw data are normalized. The visual data used in the experiment as shown in Fig. 6.

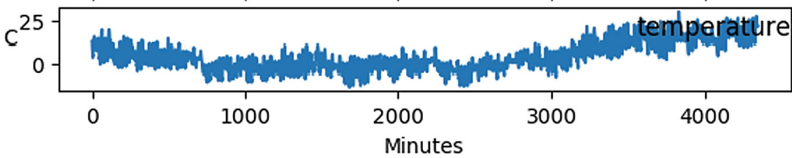
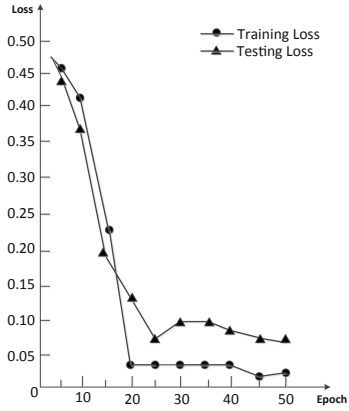


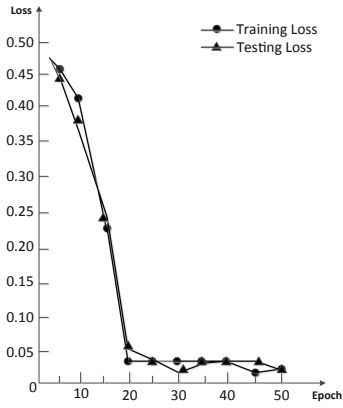
Fig. 6. Data visualization

5.2 Result and Analysis

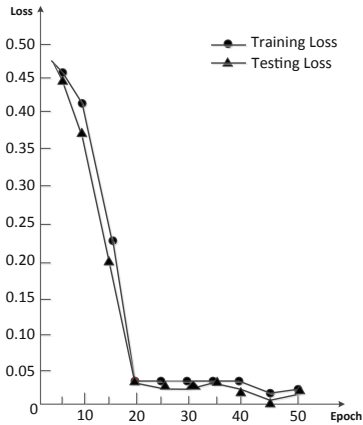
The experiment first needs to determine the batch size of the model. Figure 7 shows the change in the loss of the model for different batch sizes under the same parameters. It can be seen that the data sets corresponding to three different Batch_sizes have no excessive fluctuations; when Batch_size=64 and Batch_size = 128, the final loss is small; but when Batch_size = 128, there is a test loss value which always appear below the training loss value. Therefore, this paper selects Batch_size = 64 to train the model.



(a) Batch_size=32



(b) Batch_size=64



(c) Batch_size=128

Fig. 7. Comparison of loss in batch sizes

Then, in order to compare the performance of the model in different deep learning algorithm, we divide the above dataset into 40 training sets, each of which contains series of environmental temperature data (i.e. 1–15 h, 2–16 h, ..., 10–24 h,, 1–30 h, 2–31 h, ..., 10–39 h), and the sample data of the following hours is used as the testing set to process model accuracy test.

We calculate the error score of the model using the predicted value and the actual value of the testing set and generate the Root Mean Square Error (RMSE) for error comparison as follows, where y is an actual value and \bar{y} is a prediction value.

$$RMSE = \sqrt{(y - \bar{y})^2} \tag{7}$$

In order to ensure the reliability of the experiment, we performed 50 times experiments on each set of training set predictions, and calculated the mean of each group of RMSE. The experiment uses these data to predict and output prediction accuracy. Comparison results are shown in Fig. 8.

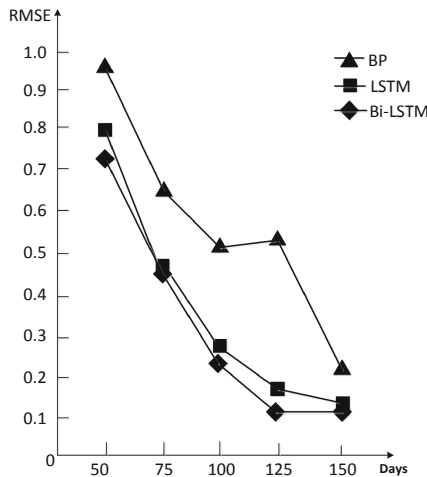


Fig. 8. Comparison of RMSE

The BP neural network, LSTM RNN are used in the comparative experiment. A three-layers BP neural network structure was consisted with an input layer, a hidden layer and an output layer. The hidden layer contains 8 hidden nodes and the layers are interconnected by correctable weight. The parameters of LSTM neural network model are the same as Bi-LSTM we used. The experiment uses these neural networks to predict the same data set and output prediction accuracy. Due to the inherent defects of the BP neural network model, which have insufficient learning of nonlinear features and the insufficiency of memory information, and it has the problem of local minimum value, which leads to the model not being able to fully training. The results show that

the BP neural network prediction result of error is higher visibly than others under the same conditions.

For three RNN, when the amount of data is large, we can get better accuracy of prediction and the differences of the prediction effects of them are not obvious. However, when the amount of data is smaller, the Bi-LSTM algorithm used in this paper can get more little error by learning and assigning the parameter features with different weights selectively compared with traditional LSTM and BP algorithm.

Next, for comparing the network bandwidth load between our proposed method with Sensor-Cloud method above mentioned reasonably, we using the same server system as they do to run the test experiment. The sensor data is divided into six group and the nload tool is responsible for monitoring the bandwidth load during data collection. The comparison result is shown in Fig. 9, where Sensor-Cloud represents the delay time from the device sensor to the cloud server, the Sensor-EdgeDL-CloudDL denotes the delay time from device sensor to the edge node. The result shows that the method of transmitting sensor data to the cloud directly (Sensor-Cloud mode) has the largest bandwidth load. Using deep learning technology to train and predict future data in the cloud can reduce the collection of sensing data than the Sensor-Cloud mode when the prediction error is lower than the set threshold. Using the method proposed in this paper, the prediction model is deployed to the edge node and the data comparison work at the edge layer can further reduce the packet collection amount and achieve the minimal network load.

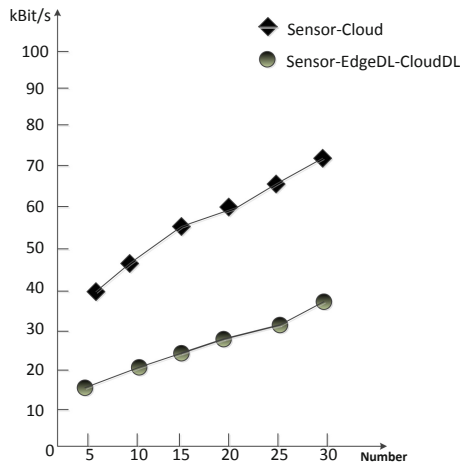


Fig. 9. Comparison of bandwidth load

In addition to the above experiment, the end-to-end delay time was also evaluated and measured. Each measurement was made 10 times and the results of the evaluations can be seen in Table 1, where Sensor-Cloud represents the delay time from the device sensor to the cloud server, Sensor-Edge denotes the delay time from device sensor to the edge node, Number is the number of data packet collection. The sensor data is

divided into five groups and each group includes 100 sensing data, the results show that the delay time of Sensor-Cloud is longer than the delay time of Sensor-EdgeDL-CloudDL when the network and devices are in the same situation.

Table 1. Delay measurement result

Number	Method	
	Sensor-cloud	Sensor-edge
100	0.14 s	0.08 s
200	0.62 s	0.51 s
300	1.10 s	0.93 s
400	1.35 s	1.24 s
500	1.87 s	1.77 s

6 Conclusion

In this paper, an IoT data collection method based on cloud-edge coordination is presented. Compared with the original method of IoT data access and collection architecture, we proposed a data collection method that combines cloud with edge nodes by using deep learning technology to provide data collection services. The cloud is responsible for storing the large amount of historical sensor data, training the deep learning model, and deploying the model to the edge side. The edge node will receive the model of data prediction and real-time data comparison and then determine whether the real data will be uploaded to the cloud and optimize the model based on the prediction accuracy. Experiments show that the method we proposed of combining edge with cloud can not only increase the speed of computing services and satisfy the real-time requirements of data collection, but also reduces the network traffic and eliminates bandwidth restrictions effectively.

It is worth mentioning that the presented model and testbed system are still under development. Moreover, aiming at the problem of how to conserve energy, we plan to add a distributed learning model on the sensor device and simulate the data stream in the edge, instead of transmitting all raw sensor values to the edge. These are topics for future work.

Acknowledgement. This work is supported by National Key R&D Plan (No: 2018YFB1402500); National Natural Science Foundation of China (No. 61832004).

References

1. Shi, W., Cao, J., et al.: Edge computing: vision and challenges. *IEEE Internet Things J.* **3**(5), 637–646 (2016)
2. Yang, Y., Li, K., Xu, H.D., et al.: Fog computing-enabled robot simultaneous localization and mapping. *Chin. J. Internet Things* **2**(2), 33–40 (2018)

3. Madria, S.K.: Sensor cloud: a cloud of sensor networks. In: IEEE International Conference on Cloud Engineering Workshop. IEEE (2016)
4. Truong, T., Dinh, A., Wahid, K.: An IoT environmental data collection system for fungal detection in crop fields. In: Electrical and Computer Engineering, pp. 1–4. IEEE (2017)
5. Peralta, G., Iglesiasurkia, M., Barcelo, M., Gomez, R., Moran, A., Bilbao, J.: Fog computing based efficient IoT scheme for the Industry 4.0. In: Electronics, Control, Measurement, Signals and Their Application to Mechatronics, pp. 1–6. IEEE (2017)
6. Oma, R., Nakamura, S., Enokido, T., Takizawa, M.: An energy-efficient model of fog and device nodes in IoT. In: Proceedings of 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 301–306 (2018)
7. Hinton, G.E., Osindero, S., Teh, Y.W.A.: Fast learning algorithm for deep belief nets. *Neural Comput.* **18**(7), 1527–1554 (2006)
8. Xiaoyun, Q., Xiaoning, K., Chao, Z., Shuai, J., Xiuda, M.: Short-term prediction of wind power based on deep long short-term memory. In Proceedings of the 2016 IEEE PES Asia-Pacific IEEE, Power and Energy Engineering Conference (APPEEC), Xi'an, China, 25–28 October 2016, pp. 1148–1152 (2016)
9. Zhang, W.: Design and Implementation of Intelligent Home System Based on Machine Learning. Jilin University (2016)
10. Xin, W., Ji, W., Chao, L.: Fault time series prediction based on LSTM recurrent neural network. *J. Beijing Univ. Aeronaut. Astronaut.* **44**(4), 772–784 (2018)
11. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **18**(5), 602–610 (2005)
12. Zhao, R., Yan, R., Wang, J., Mao, K.: Learning to monitor machine health with convolutional bi-directional LSTM networks. *Sensors* **17**(2), 273 (2017)
13. Deepa, A., Dharani, R., Kalaivani, S., Parkavi, P.M.: Live video streaming system using Raspberry pi with cloud server. *IJAICT* **2**(11), 1075–1077 (2016)
14. Ismail, A.A., Hamza, H.S., Kotb, A.M.: Performance Evaluation of Open Source IoT Platforms. In: 2018 IEEE Global Conference on Internet of Things (GCIoT), Alexandria, Egypt, pp. 1–5 (2018)



A Malicious Anchor Detection Algorithm Based on Isolation Forest and Sequential Probability Ratio Testing (SPRT)

Jun Peng¹ and Xingcheng Liu^{1,2}(✉)

¹ School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China

² School of Information Science, Xinhua College of Sun Yat-sen University, Guangzhou 510520, China
isslxc@mail.sysu.edu.cn

Abstract. Many applications ask for information of nodes in wireless sensor networks (WSNs), among which the node location is an important type of information. In WSNs, localization of target node is often obtained with aid of anchors with providing distance-related information in many algorithms. However some anchors may be attacked in real environments. In order to guarantee the accuracy of localization, it is necessary to identify the malicious anchors under attack. In this paper, we propose a localization detection algorithm with malicious anchors existing in WSNs. The algorithm firstly utilizes Isolation Forest to confirm the reference anchors. After obtaining the initial position estimate of the target nodes, we establish a detection model using the consistency of the measuring distance and the Euclidean distance to the initial position estimate. Finally sequential probability ratio testing (SPRT) is carried out on the remained anchors. The simulation results demonstrate the proposed algorithm can efficiently identify the malicious anchors and outperforms other existing algorithms.

Keywords: Wireless sensor networks (WSNs) · Malicious anchor detection · Isolation Forest · Sequential probability ratio testing (SPRT) · Secure localization

1 Introduction

Wireless sensor networks (WSNs) contain a large amount of wireless sensor nodes deployed over specific area. The sensor nodes have the capacity of data processing, information collecting and etc. WSNs can gather and process the information

Supported by the National Natural Science Foundation of China under Grant 61572534 and Grant 61873290, by the Special Project of Promoting Economic Development in Guangdong Province under Grant GDME-2018D004, and by the Opening Project of Guangdong Province Key Laboratory of Information Security Technology under Grant 2017B030314131.

about the monitored object in the coverage area and send to the observer [1]. As soon as WSNs appears, it has been widely concerned in many fields [2], and was applied to environmental monitoring, medical health, forest fire prevention, National defense military and many other fields. Node location information is crucial for these applications, so node localization in WSNs is a significant issue.

The node localization algorithms are usually divided into two categories: range-based [3] and range-free [4, 5]. The range-based algorithms utilize some physical measuring techniques to obtain the range between anchors and the target nodes such as Time of Arrival (ToA), Time-Difference of Arrival (TDoA), Angle of Arrival (AoA), and Received Signal Strength Indication (RSSI). The range-free algorithms such as DV-Hop [6] locate the target nodes via the connectivity of the network.

WSNs are often deployed on the unattended area like forest and marsh. Therefore, the anchors may be vulnerable on account of the environmental implication. Furthermore, anchors may be compromised by enemy when WSNs are used in military. All of the above situation will make some of the anchors under attack. Anchors under attack are considered as malicious in this paper. We focus on the secure localization of range-based algorithms in presence of malicious anchors.

Many malicious anchors detection works in WSNs have been done before. Method proposed in [7] filters out malicious anchor signals on the basis of the consistency among multiple anchor signals. Gradient descent (GD) method with a selective pruning stage for inconsistent measurements is used to achieve localization [8]. Sparse recovery formulation was used to solve the secure localization problem in the algorithm proposed in [9]. Using clustering and consistency of the RSSI and ToA measurements can achieve good detection of malicious anchors [10].

In this paper, we propose a malicious detection algorithm based on Isolation Forest and Sequential Probability Ratio Testing (SPRT) to address the secure localization problem of range-based algorithms.

The rest of the paper are organized as follows: Sect. 2 introduces the network model and problem formulation. Then the proposed malicious anchor detection algorithm is described in details in Sect. 3. Our simulation results and comparison are demonstrated in Sect. 4. The last section is the summary and conclusion of this paper.

2 Network Model and Problem Formulations

2.1 Network Model

The network model is set up as a two-dimensional wireless sensor network (WSN). We assume that the network is stable. In the situation of our test, the measurements of the distance between anchors and target node are available. Since ToA technology requires strict synchronization while TDoA does not have such a requirement, we utilize TDoA technology to obtain the range measurements in the proposed algorithm.

The notations used in this paper are listed in Table 1.

Table 1. Summary of notations

Notations	Meanings
n	Number of anchors
m	Number of malicious anchors
c	Number of reference anchors
k	Number of observations
\mathbf{A}_i	Location of the i -th anchor
\mathbf{T}	Location of the target node
\mathbf{T}_f	Initial location estimate of the target node
d_i	Measured distance of the i -th anchor
n_i	Noise components of the i -th anchor
u_i	Attack components of the i -th anchor
σ	Std deviation of noise components
μ_δ	Mean value of attack components
σ_δ	Std deviation of attack components
p_0	Null hypothesis threshold
p_1	Alternative hypothesis threshold
α	False negative rate
β	False positive rate

2.2 Problem Formulation

One target node \mathbf{T} which need to be located and n anchors whose locations are already known as $\{\mathbf{A}_1, \dots, \mathbf{A}_n\}$ are randomly deployed in the network. Among all the anchors, m of them are malicious. We assume that all the anchors are in the communication range of the target node. The true distance between i th anchor and the target node is $\|\mathbf{A}_i - \mathbf{T}\|$. Consider noise of measurement and attack of the malicious anchors, a model of distance measurement between i -th anchor and the target node d_i can be established as Eq. (1):

$$d_i = \begin{cases} \|\mathbf{A}_i - \mathbf{T}\| + n_i, & \text{if } i\text{-th anchor is honest} \\ \|\mathbf{A}_i - \mathbf{T}\| + n_i + u_i, & \text{if } i\text{-th anchor is malicious} \end{cases} \quad (1)$$

where, the noise components in the measurements of the i -th anchor and the target node n_i are assumed to be the independently distributed white Gaussian variables, given by $n_i \sim \mathcal{N}(0, \sigma^2)$. The attack components $u_i \sim \mathcal{N}(\mu_\delta, \sigma_\delta^2)$. In the actual cases, the ranging impact of malicious anchors is usually characterized by increasing measurements of distance. Therefore, in this paper, we set $\mu_\delta > 0$.

3 Proposed Algorithm

In this section, we propose a malicious anchors detection algorithm using Isolation Forest and Sequential Probability Ratio Testing (SPRT) to achieve secure localization in WSNs. The proposed algorithm can be divided into the following three steps:

1. Determine the reference anchors by Isolation Forest algorithm.
2. Calculate the reference error interval and establish the testing model.
3. Carry out Sequential Probability Ratio Testing (SPRT) on the remained anchors.

The flow chat of the proposed algorithm is presented in Fig. 1. More specific description is as follows.

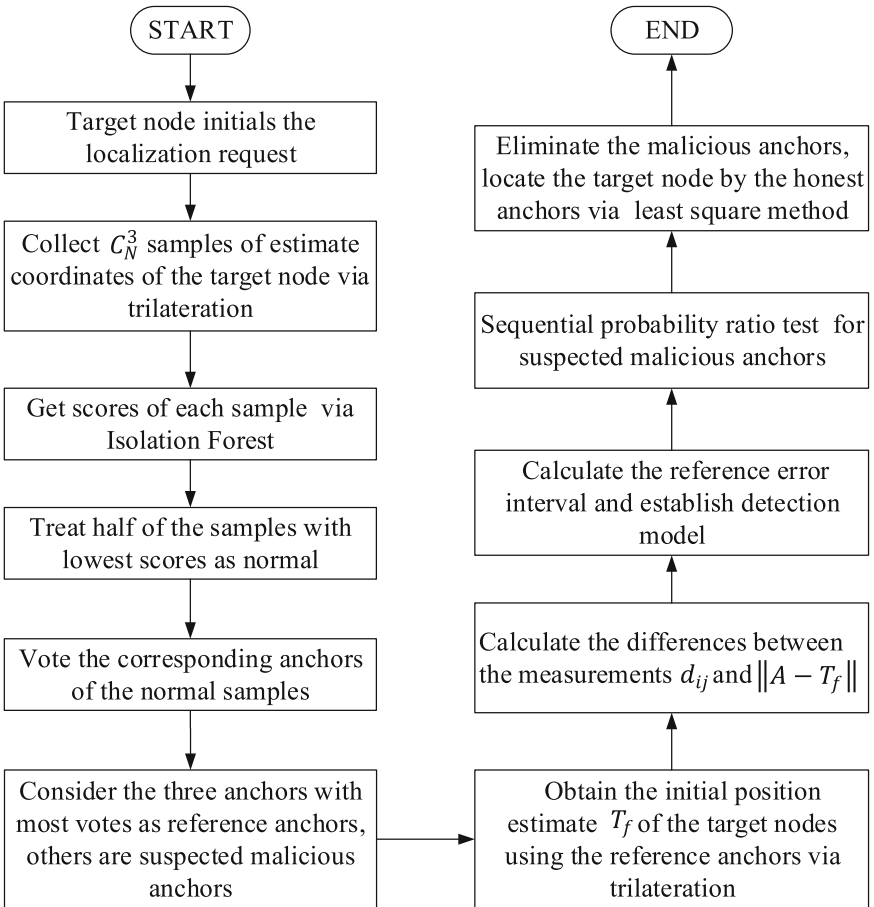


Fig. 1. Flow chart of the proposed algorithm

3.1 Determine the Reference Anchors

The target node sends signal for request of localization, then every anchor within the range of the target node transmits the packet that consists of self-location, identification number and the distance measurement from itself to the target node. If the number of anchors are larger than 3, i.e. $n > 3$, select information provided by any three of all the anchors to locate the target node via trilateration. The total number of the estimate coordinates of the target node is C_n^3 . Record the identification numbers of corresponding anchors of each estimate coordinate. The estimate coordinates, which were supplied only from the honest anchors, are considered as normal samples. As long as any of the three anchors is malicious, the estimate coordinates are considered to be abnormal. Normal samples are only affected by measurement noise and trilateration errors while the abnormal samples affected by attack in addition. Thus, normal samples are close to the real location of the target node and densely distributed, while the abnormal samples are relatively far away and sparsely distributed.

Isolation Forest is an effective outlier detection algorithm [11]. Outliers refer to data points that are sparsely distributed and far away from high density areas. Combined with this feature, the technique can be used to deal with the estimate coordinates. Isolation Forest can score each sample. The score represents the abnormal degree of the samples, higher scores correspond to higher abnormal degree. In our algorithm, we consider half samples with lowest scores are normal. Vote the identification numbers of anchors corresponding to the samples that is judged to be normal by Isolation Forest. One time of occurrence plus one vote. Three anchors with highest votes are identified as the reference anchors used in the next step. The reference anchors are considered to be honest.

3.2 Establish the Testing Model

Observe the distance between all anchors and the target node for k times, d_{ij} denotes the j -th measurement of i th anchor. Self-locations and mean values of the k measurements of the reference anchors are utilized to obtain the initial estimate location \mathbf{T}_f of the target nodes via trilateration. Euclidean distance between \mathbf{T}_f and i th anchor is $\|\mathbf{A}_i - \mathbf{T}_f\|$, difference between the j -th measurement and Euclidean distance of i -th anchor D_{ij} is defined as Eq. (2):

$$D_{ij} = |d_{ij} - \|\mathbf{A}_i - \mathbf{T}_f\||. \quad (2)$$

If i -th anchor is honest, only measurement noise and trilateration error may affect D_{ij} , and the value of D_{ij} is within an acceptable range. However, if i -th anchor is malicious, due to the impact of attack, D_{ij} may be out of the range. The range can be obtained using the information provided by the reference anchors.

A reference anchor is considered as an individual and each D_{ij} is considered as a sample. Each individual contains k samples. The mean value and variance of the i -th individual are calculated by Eqs. (3) and (4).

$$\bar{D}_i = \frac{\sum_{j=1}^k D_{ij}}{k}, \quad (3)$$

$$s_i^2 = \frac{\sum_{j=1}^k (D_{ij} - \bar{D}_i)^2}{k - 1}. \quad (4)$$

The mean value \bar{D} of all the individuals is show in Eq. (5), variability of the distribution of individual mean \bar{D}_i estimate by Eq. (6):

$$\bar{D} = \sum_{i=1}^c \frac{\bar{D}_i}{c}, \quad (5)$$

$$s_d^2 = \sum_{i=1}^c \frac{(\bar{D}_i - \bar{D})^2}{c - 1}. \quad (6)$$

In which, c represents the number of individuals. In the proposed algorithm, the number of reference anchors is 3, i.e. $c = 3$. With reference to the variance of each individual, the variance of all the reference anchors can be estimated by Eq. (7).

$$s_a^2 = \sum_{i=1}^c \frac{k - 1}{N - c} s_i^2 \quad (7)$$

where $N = c \times k$. The general variance of all the samples are shown in Eq. (8).

$$s_g^2 = s_d^2 + \left(1 - \frac{1}{m_h}\right) s_a^2 \quad (8)$$

where, m_h is Harmonic Mean (HM) of time of observations, because we observe k times for every anchor, so we have $m_h = k$.

The reference error interval $[D_{min}, D_{max}]$ can be given as Eq. (9),

$$D_{min} = \bar{D} - (z_{1-\frac{\alpha}{2}}) \times s_g, \quad (9)$$

$$D_{max} = \bar{D} + (z_{1-\frac{\alpha}{2}}) \times s_g. \quad (10)$$

$z_{1-\frac{\alpha}{2}}$ refers to the upper quartile of $(1 - \frac{\alpha}{2})$ in standard normal distribution, where α represents the significance level.

3.3 Sequential Probability Ratio Testing (SPRT)

Sequential Probability Ratio Testing (SPRT) [12] belongs to hypothesis testing. In this paper, we utilize the technique to detect the malicious anchors because it can reduce the times of testing and has high reliability.

Before testing, we establish a Bernoulli random variable Z_{ij} ,

$$Z_{ij} = \begin{cases} 0, & D_{min} < D_{ij} < D_{max} \\ 1, & \text{others} \end{cases} \quad (11)$$

The probability p of Bernoulli distribution is defined as $p = P(Z_{ij} = 1)$, and $P(Z_{ij} = 0) = 1 - P(Z_{ij} = 1)$. If p is bigger than the predefined threshold p_t , the i th anchors are considered to be malicious. However the p_t is hard to define. In order to reduce errors caused by SPRT, two limits p_0 and p_1 can be introduced. We introduce two hypotheses:

1. H_0 : $p < p_0$, the anchor is honest,
2. H_1 : $p > p_1$, the anchor is malicious.

User-configured false positive rate (FPR) is defined as α , and false negative rate (FNR) is β . The upper bounds are provided by α and β . We consider each observation independent, so each Z_{ij} is independent. Define l_{ij} as the log-probability ratio on j samples of i th anchor, given as Eq. (12),

$$l_{ij} = \frac{P(Z_{i1}, \dots, Z_{ij}|H_1)}{P(Z_{i1}, \dots, Z_{ij}|H_0)} = \sum_{q=1}^j \frac{P(Z_{iq}|H_1)}{P(P(Z_{iq}|H_0))}. \quad (12)$$

Let S_{ij} be the times $Z_{ij} = 1$ of j samples, then we have

$$\ln S_{ij} = S_{ij} \ln\left(\frac{p_1}{p_0}\right) + (j - S_{ij}) \ln\left(\frac{1 - p_1}{1 - p_0}\right). \quad (13)$$

According to the log-probability ratio l_{ij} , the following results can be inferred:

1. $S_{ij} \leq L_j$, accept H_0 , terminate the test.
2. $S_{ij} \geq U_j$, accept H_1 , terminate the test.
3. $L_j < S_{ij} < U_j$, continues the test with another observation.

L_j denotes the acceptable number of samples that out of the reference error interval, while U_j is corresponding unacceptable number of all j samples.

$$L_j = \frac{\ln \frac{\beta}{1-\alpha} + j \ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1}{p_0} - \ln \frac{1-p_1}{1-p_0}}, \quad (14)$$

$$U_j = \frac{\ln \frac{1-\beta}{\alpha} + j \ln \frac{1-p_0}{1-p_1}}{\ln \frac{p_1}{p_0} - \ln \frac{1-p_1}{1-p_0}}. \quad (15)$$

4 Simulation Results

To evaluate the performance of the secure localization algorithm for wireless sensor networks proposed in this paper, we utilize two types of evaluation criteria: True Positive Rate (TPR), which denotes the proportion of detected malicious anchors to the total number of malicious anchors, and False Positive Rate (TPR), which denotes the proportion of honest anchors that are misjudged as malicious. All the simulations were performed on *Matlab2016a*.

At first, we consider the situation that there is only one malicious anchors while the total number of anchors N is relatively small. MNDC and its improved version EMDC were proposed to achieve malicious anchor detection and secure localization in such a situation [10]. There is a very important premise condition to implement the method in MNDC and EMDC: to guarantee the measurements of RSSI are not attacked while the measurements of ToA are under attack. However attack may affect all of the measurements in practical. But such a

Table 2. Parameters setting of Experiment 1

n	m	k	σ	μ_δ	p_0	p_1	α	β
6	1	30	2 m	12 m	0.1	0.9	0.01	0.01

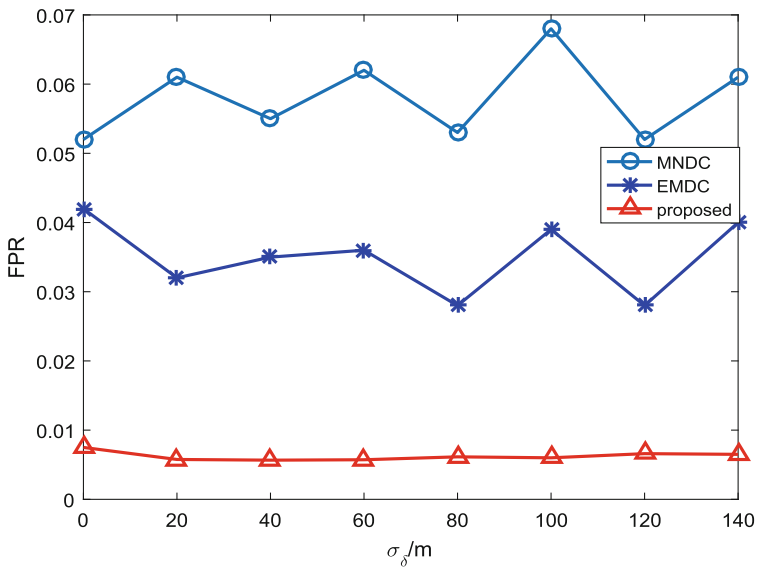
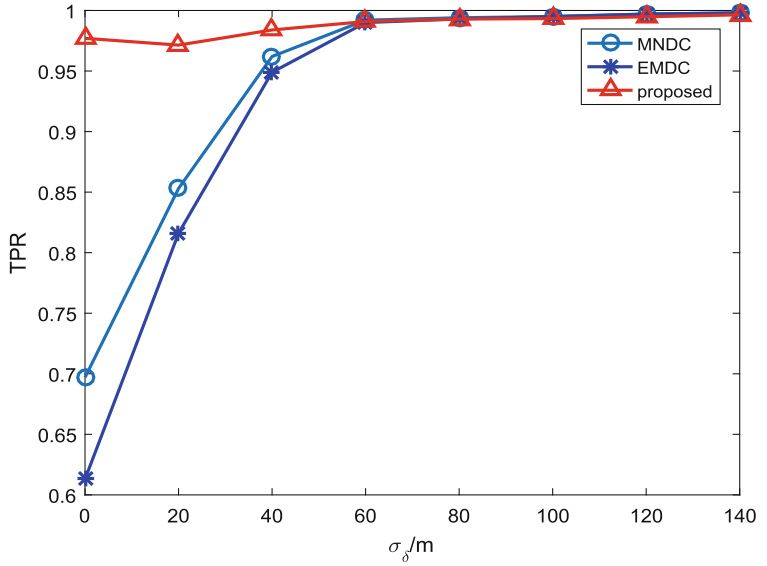


Fig. 2. Performance with varying σ_δ

premise condition is not necessary in our algorithm. With reference to [10], nodes are deployed in a $100\text{ m} \times 100\text{ m}$ square target field, and the parameters are set as Table 2.

The TPR and FPR curves vary the standard deviation of the attack σ_δ changes of the algorithms are shown in Fig. 2. As the σ_δ increases, the detection performance of the proposed method increases. That is because the detection effect of Isolation Forest becomes better and the number of samples out of the range of reference error interval increases if the σ_δ increases. Our algorithm maintains TPR higher than 0.9 while FPR smaller than 0.01. Compare to MNDC and EMDC, we find our algorithm outperforms at both TPR and FPR, in particular when the standard deviation lower than 40m. In combination with the condition that we do not need to guarantee the measurements of RSSI correct all the time, we could draw a conclusion that the proposed algorithm outperforms both MNDC and EMDC.

Then, we consider a situation where there is more than one malicious anchor in the system. Ravi Garg et al. proposed two kinds of GD algorithms, one of which is the fixed step size algorithm GD_f , the other is the variable step size

Table 3. Parameters setting of Experiment 2

n	m	k	σ	μ_δ	p_0	p_1	α	β
30	9	30	2 m	4 m	0.1	0.9	0.01	0.01

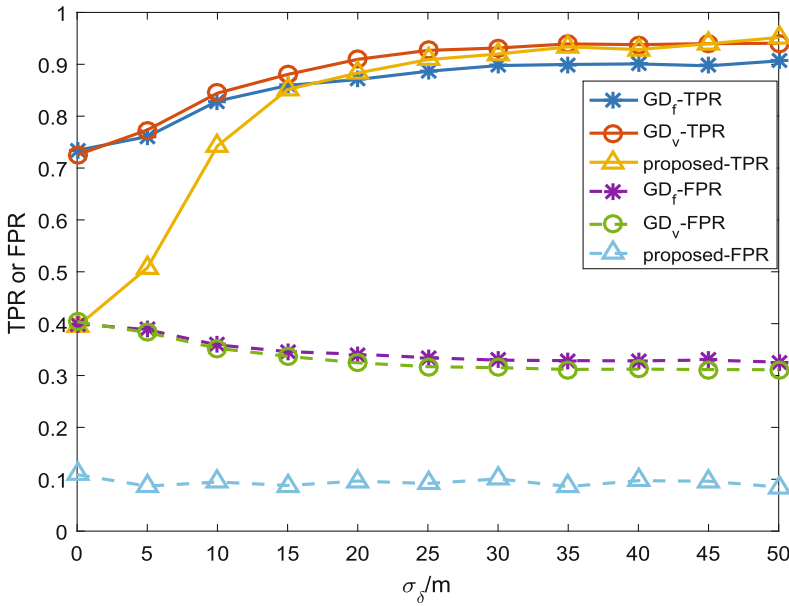


Fig. 3. TPR and FPR curves with varying σ_δ

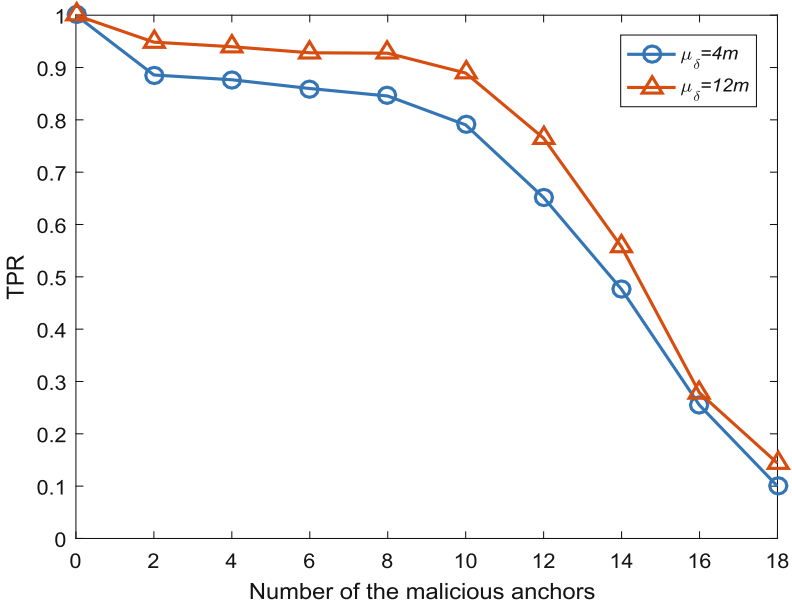
Table 4. Parameters setting of Experiment 3

n	k	σ	σ_δ	p_0	p_1	α	β
30	30	2 m	12 m	0.1	0.9	0.01	0.01

algorithm GD_v [8]. For GD_f , the step size is fixed at 0.5 m, and for GD_v , the rule of the change of the step size is $\gamma(i) = 15 - \frac{15(i-1)}{M}$, in which $\gamma(i)$ represents the step size of the i -th iteration, and M is the maximum number of iterations. For both GD_f and GD_v , the gradient threshold to switch to the selecting stage is fixed at 0.9 while M is set to 500. Half of the force vectors are pruned. Nodes are deployed in a $60\text{ m} \times 60\text{ m}$ square target field, and other parameters are set as Table 3.

The TPR and FPR curves of the proposed algorithm and GD algorithms are shown in Fig. 3. As the attack intensity increases, the TPR and FPR of GD algorithms maintain relatively high. The main reason of this phenomenon is that the percentage of the pruned vector is fixed at 50%. The TPR of the proposed algorithm is close to GD algorithms when σ_δ exceeds 15 m. An outstanding advantage of the proposed algorithm is that the FPR maintains around 0.1, which greatly outperforms the GD algorithms.

Finally, we observe the TPR curves with varying number of malicious anchors under different mean value of attack μ_δ . Set the parameters as Table 4.

**Fig. 4.** TPR curves with varying number of malicious anchors under different μ_δ

As is seen in Fig. 4 TPR drops as the number of anchors increases and rises as μ_δ increases. As the percentage of the malicious anchors is smaller than 30%, TPR is greater than 0.9 when $\mu_\delta = 12$ m, and greater than 0.8 when $\mu_\delta = 4$ m. When the proportion of malicious anchors exceeds 30%, TPR drops faster.

5 Conclusions

A secure range-based localization algorithm has been introduced in this paper to detect the malicious anchors. We utilize Isolation Forest to determine reference anchors and the initial estimated position of the target nodes. Then a testing model can be established using the difference between the measurements of distances and the Euclidean distance from the reference anchors to the initial estimated position. Finally Sequential probability ratio testing (SPRT) is carried out on the remained malicious anchors. Our simulation results demonstrate that the proposed algorithm outperforms other state-of-art algorithms. The future work direction would be to improve the detection rate when the proportion of malicious anchors increases.

References

1. Akyildiz, I.F., et al.: A survey on sensor networks. *Commun. Mag.* **40**(8), 102–114 (2002)
2. Liu, X., et al.: An optimization scheme of enhanced adaptive dynamic energy consumption based on joint network-channel coding in WSNs. *Sensors J.* **17**(18), 6119–6128 (2017)
3. Liu, X., et al.: Range-based localization for sparse 3-D sensor networks. *IoT-J.* **6**(1), 753–764 (2019)
4. Lee, J., et al.: Novel range-free localization based on multidimensional support vector regression trained in the primal space. *IEEE Trans. Neural Netw. Learn. Syst.* **24**(7), 1099–1113 (2013)
5. Assaf, A.E., et al.: Robust ANNs-based WSN localization in the presence of anisotropic signal attenuation. *Wirel. Commun. Lett.* **5**(5), 504–507 (2016)
6. Niculescu, D., et al.: Ad hoc positioning system (APS). In: *Global Telecommunications Conference*, pp. 2926–2931. IEEE (2001)
7. Liu, D., et al.: Attack-resistant location estimation in sensor networks. *ACM Trans. Inf. Syst. Secur.* **11**(4), 1–39 (2008)
8. Garg, R., et al.: An efficient gradient descent approach to secure localization in resource constrained wireless sensor networks. *IEEE Trans. Inf. Forensics Secur.* **7**(2), 717–730 (2012)
9. Zhang, Q., et al.: Sparse recovery formulation for secure distance-based localization in the presence of cheating anchors. *Wirel. Netw.* **24**(7), 2657–2668 (2018)
10. Liu, X., et al.: A range-based secure localization algorithm for wireless sensor networks. *Sensors J.* **19**(2), 785–796 (2018)
11. Liu, F.T., et al.: Isolation forest. In: *8th International Conference on Data Mining*, pp. 413–422. IEEE (2008)
12. Ho, J.W., et al.: Zone trust: fast zone-based node compromise detection and revocation in wireless sensor networks using sequential hypothesis testing. *IEEE Trans. Dependable Secur. Comput.* **9**(4), 494–511 (2012)



Noisy Data Gathering in Wireless Sensor Networks via Compressed Sensing and Cross Validation

Xiaoxia Song^(✉), Yong Li, and Wenmei Nie

College of Computer & Network Engineering, Shanxi Datong University,
Datong 037009, China
sxxly2002@163.com

Abstract. In wireless sensor networks (WSNs), sensor data are usually corrupted by the noise. Meanwhile, it is inevitable to face the problems of node energy in WSNs. For both of these questions, this paper proposes a data gathering method via compressed sensing combined with cross validation. In the proposed method, data gathering via CS can save and balance energy consumption of sensor nodes due to the features of CS, and CV technique is used to judge whether stable reconstruction have been obtained. This method is essentially an adaptive intelligent method. Unlike the existing methods, the proposed method does not need the knowledge of signal sparsity, noise information and/or regularization parameter while those knowledge is expensive to acquire, especially in adaptive systems. That is to say, the method proposed in this paper is not sensitive to signal sparsity, noise, regularization parameters and/or other information when it is used for WSNs data collection for noise case, but the existing methods rely heavily on the prior information. Experimental results show that the proposed data gathering method can obtain stable reconstruction results for noisy WSNs in the case of unknown signal sparsity, noise and/or regularization parameters.

Keywords: Data gathering · Wireless sensor networks · Compressed sensing · Cross validation

1 Introduction

Wireless sensor networks (WSNs) are composed of a large number of low-cost micro sensor nodes deployed in a monitoring area. They often take the form of multi-hop self-organizing networks to sample, sense and process the information of the objects to be sensed [3]. So far, WSNs have been applied in a wide range of applications scenarios since it can change the living and working ways of the human being, such as military, environmental monitoring, transportation, agriculture, medical and health for their application advantages [1, 4, 7]. However, energy problem of sensor nodes in WSNs is one of main challenges for its application since each sensor node is usually powered by batteries which will not be convenient to be charged in many cases.

It is an important way to reduce energy consumption in data gathering of WSNs by different organizing ways, gathering technology and intelligent methods [10–12, 17].

The articles [10, 12] realize data gathering in WSNs by organizing the sensor nodes into hierarchical and tree structures. The cluster-based data gathering methods proposed in [11, 17] usually bring to the early death of some sensor nodes and further lead to unbalanced energy consumption. Fortunately, data gathering methods based on compressed sensing (CS) can save and balance energy consumption of sensor nodes since it can capture the feature of a k -sparse signal by a small number of measurements [14]. CS, proposed by Candes and Donoho in 2006 [5, 8], indicates that a small number of linear measurements of the noiseless signals can be used to perfectly reconstruct the signal with high probability. CS is widely concerned in some areas, such as signal processing, applied mathematics, and statistics and wireless sensor networks [16, 19]. In the noiseless setting, there exists an experience that a certain number of measurements can be used to reconstruct perfectly the signal when restricted isometry property (RIP) is satisfied and the sparsity of the signal is known.

The performance of plain CS drops dramatically when the measurements are corrupted by the noise. Therefore, some researches also provide some theory results [2] for the problem. Donoho and Candes show that for $M = O(k \log(N/k))$ samples there are efficient algorithms, for example quadratic programming algorithms, to provide stable reconstruction results. Some articles illustrate that the sparse signals can be stably recovered from noisy measurements under the mutual incoherence property (MIP) framework. And the other articles show that $M = O(k)$ measurements are necessary and sufficient to reconstruct stably the signals.

The methods above can be used to achieve the sensor data in the sink node of WSNs when sensor data are corrupted by the noise. But these methods need some knowledge of the signal and/or noise [6, 9, 13, 14, 20]. Some methods, such as [13, 20], need to know the variance of the noise because their iterative algorithms use the information of the noise magnitude to establish a stopping criterion. The other methods, such as basis pursuit denoising [6], lasso [15], and gradient projection [9], need to estimate the regularization parameter, which depends on two unknown factors including the noise level and signal sparsity. Unfortunately, they need to estimate these parameters at great cost, especially in adaptive systems. Meanwhile, the effectiveness of these methods relies heavily on the accuracy of these assessment parameters.

In this paper, we propose a noisy data gathering method for wireless sensor networks via compressed sensing combined with cross validation. In the proposed method, data gathering via CS can save and balance energy consumption of sensor nodes in WSNs. To obtain the stable recovery in the sink, cross validation (CV) [18] is introduced into CS reconstruction algorithm. So the proposed method can be referred to as CSCV. In fact, it is an improved CS data gathering method for noisy WSNs. Unlike the existing methods, it does not need the knowledge of signal sparsity and/or noise. Before acquiring initial measurements in the proposed method, some additional measurements with $\log_2 N$ components are acquired as the estimation set. After acquiring initial measurements, CV technique is used to judge whether these measurements can stably reconstruct the signal or not. If those measurements cannot be used to reconstruct sensor data, some new measurements are acquired step by step until the terminal condition is satisfied. Experimental results show that the proposed method can obtain stable reconstruction results for noisy WSNs in the case of unknown signal sparsity, noise and/or regularization parameters.

Before proceeding, we define some denotations. Let \mathbf{x} be the original signal with the length N and the sparsity k . The vector $\mathbf{y} = [y_1, y_2, \dots, y_M]$ represents noisy measurements, i.e., M is the measurement numbers. Then $\mathbf{y} = \Phi\mathbf{x} + \mathbf{m}$, where Φ is the random measurement matrix, \mathbf{m} is Gaussian noise. Let $\mathbf{w} = [w_1, w_2, \dots, w_P]$ ($P = \log_2 N$) be estimation set, and $\mathbf{w} = \Psi\mathbf{x} + \mathbf{m}$. Suppose $\hat{\mathbf{x}}_M$ represent the reconstruct signal by taking M noisy measurements. The denotation τ represents a given reconstruction precision.

The remainder of the paper is structured as follows: in Sect. 2, methods and results of previous studies on related work are summarized. In Sect. 3, a detailed introduction is presented to explain the framework of the proposed method, the model of data gathering in WSNs via CS, the judgment method by CV, and CSCV data gathering method. In Sect. 4, evaluation metrics, the experimental setup and results are presented. In Sect. 5, the conclusions are given.

2 Related Work

In this section, we firstly introduce CS for noiseless and noisy signals, and illustrate CS data gathering in WSNs. Then, the difficult issue how to determine the stable recovery in sink node is analyzed. Finally, CV technique is introduced to solve this difficult issue.

2.1 Compressed Sensing

CS is a new information acquirement theory, which is proposed by Candes and Donoho in 2006 [5, 8]. The main idea of CS is given as follows. Suppose the coefficients of a signal \mathbf{x} with the length N is sparse or compressible in certain a orthogonal basic or tight framework Ψ . If these coefficients are projected to a measurement matrix Φ with the dimensions $M \times N$ ($M \ll N$), which is incoherent with Ψ , the observations \mathbf{y} can be obtained. Then \mathbf{x} can be achieved in the sink by the following two steps. (i) Achieve Θ by solving the model (1). (ii) Compute \mathbf{x} by $\mathbf{x} = \Psi\Theta$.

$$\min\|\Theta\|_1 \quad s.t. \quad \Phi\Psi\Theta = \mathbf{y} \quad . \tag{1}$$

Certainly if \mathbf{x} is sparse, then model (1) can be transformed to the model (2).

$$\min\|\mathbf{x}\|_1 \quad s.t. \quad \Phi\mathbf{x} = \mathbf{y} \quad . \tag{2}$$

It is known that the noiseless signal can be perfectly reconstructed by $M = O(k \log(N/k))$ (k is the sparsity of the noiseless signal \mathbf{x}) linear measurements $\mathbf{y} = \Phi\mathbf{x}$ ($\Phi \in R^{M \times N}$) ($\Phi \in R^{M \times N}$), when the matrix Φ satisfies RIP.

When the measurements are corrupted by the noise, i.e., $\mathbf{y} = \Phi\mathbf{x} + \mathbf{m}$, the noisy signal can be stably reconstructed by $M = O(k \log(N/k))$ or $M = O(k)$ noisy measurements [2]. In noise setting, the signal can be stably reconstructed when the measurements arrive at a certain numbers, while the signal cannot be perfectly

reconstructed by using even all data. So, the noise seriously affects the performance of CS. Figure 1 is given to illustrate the fact.

In Fig. 1, the horizontal-axis represents the numbers of the measurements, the vertical-axis shows the signal to noise ratio (SNR) of reconstruction signal. The measurements are corrupted by the noise with mean value 0 and variance 0.01. From this figure, we can see that the reconstruction results of 40 observations and 128 observations are basically the same. That is to say, the signal can be stably reconstructed when the measurement numbers are more than 40, while the signal cannot be perfectly reconstructed for even 128 measurements.

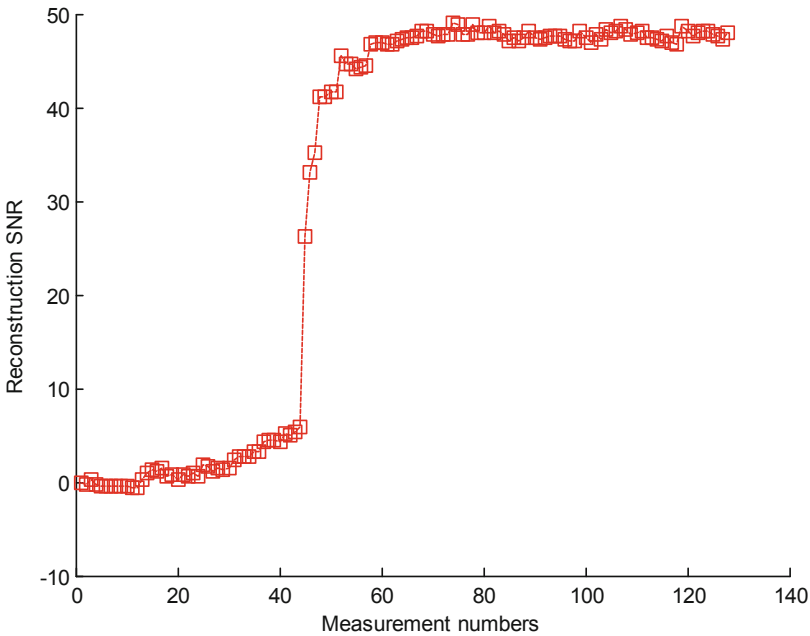


Fig. 1. Noisy reconstruction of the signal with $N = 128$, $k = 10$.

2.2 Compressed Sensing Applied in WSNs

In WSNs, each sensor usually monitors some physical parameters, such as temperature, humidity, wind speed, and soil moisture so on, of a certain area. The neighbor sensor data will be very similar since they monitor adjacent range, which is called as spatial correlation. Similarly, there is a correlation between data perceived in adjacent time on each sensor in WSNs. It's called as temporal correlation. Thus, sensor data of WSNs may be viewed as compressible signal. So it is a feasible scheme that CS is introduced into WSNs to collect the sensor data.

Some researchers have shown that data gathering methods based on CS can reduce energy consumption of sensor nodes [5, 16]. The reasons are that these methods can capture the feature of a k -sparse signal by a small number of measurements. So, the

signal can be obtained by less measurement number in the sink where the energy is readily available. When sensor data are corrupted by the noise, some methods can be used to achieve the sensor data in the sink node. However, there are not efficient solutions for data gathering in wireless sensor networks under noise cases. According to the conclusion in Fig. 1, it is feasible to collect the measurement numbers as little as possible to reconstruct the data of sensor data by CS. But it is a key issue to how to determine that a stable recovery has occurred in the sink node.

2.3 Cross Validation

CV proposed by Seymour Geisser is a statistical technique to determine the appropriate model order complexity and thus avoid overfitting a model to a set of sample data. CV's theory support is Johnson-Lindenstrauss (JL) lemma. The basic idea of it [18] is to put the original data as part of the training set, the other part as a validation set, the training set is used to train a classifier and validation set is used to test the trained model, eventually get reliable and stable model. The detailed steps are given below. (i) Divide the data set into two sets: a training/estimation set, and a test/cross validation set. (ii) The estimation set is used to obtain a model or estimation, and the cross validation set is used to validate the performance of the model or estimation. In this paper, we use CV technique to avoid overfitting of the CS noisy reconstruction algorithm.

3 The Proposed Data Gathering Method

To efficiently gather data in wireless sensor networks under noise cases, this paper propose a data gathering method via CS and CV, we call it as CSCV data gathering method. In fact, it is a modified CS combined with CV. Thus, the improved CS method is suitable for noisy data recovery. In this section, we firstly introduce the model of data gathering in WSNs via CS, and then illustrate how to determine the stable recovery in the sink node by CV, finally propose the CSCV data gathering method for WSNs.

3.1 Model of Data Gathering in WSNs via CS

In this paper, suppose the whole network is composed of N sensor nodes and a sink node. And the sensing area is a $\sqrt{N} \times \sqrt{N}$ square region, each small area is only deployed a sensor node who realizes the monitoring of an object in each square area. The sink node lies around the sensing area. Let i ($1 \leq i \leq N$) represent the i -th node of this network, x_i be the reading of the node i and indicate the monitoring feature of the i -th area. Denote $\Phi = [\Phi_1, \dots, \Phi_j, \dots, \Phi_M]^T$ ($1 \leq j \leq M$) be the measurement matrix, where Φ_j is the j -th measurement vector. The collected data of the sink node are represented as $y_j = \Phi_j \mathbf{x}$, where $\mathbf{x} = [x_1, \dots, x_i, \dots, x_N]$. When data gathering is completed, the sink node obtains the vector/measurements $\mathbf{y} = [y_1, \dots, y_j, \dots, y_M]$. Finally the sink node extracts the data of sensor nodes from \mathbf{y} by solving the model (1) or (2).

However, it is an ill-posed problem due to $M < N$. Fortunately, the above ill-posed problem can be solved by considering spatial correlation or temporal correlation of sensor data. At the same time, the model of data gathering via CS can save and balance the energy of sensor nodes to alleviate the main bottleneck of the applications of WSNs since CS can use a small number of measurements to reconstruct the high-dimension data.

3.2 The Judgment Method by CV

For the difficult problem how to determine the stable recovery in the sink node, CV is introduced to data gathering via CS. After receiving initial measurements, the homotopy algorithm is used to reconstruct the signal $\hat{\mathbf{x}}_M$. The reconstruct error is estimated by the following formula

$$err = \frac{\|\mathbb{w} - \Theta \hat{\mathbf{x}}_M\|_2}{\|\mathbb{w}\|_2} \quad (3)$$

where \mathbb{w} is the observations vector corresponding to the estimation set, Θ represents the measurement matrix corresponding to \mathbb{w} . When the formula $err \leq \tau$ holds, it means that the measurements may be enough to reconstruct stably the signal, otherwise it implies that the measurements are not enough to stably reconstruct the signal, and it needs to further acquire some measurements until the terminal condition is satisfied.

3.3 CSCV Data Gathering Method

In this section, for the situation that sensor data are polluted by noise, we propose a data gathering method for WSNs via CS and CV. In the proposed method, CS is used to save and balance energy consumption of the sensor node in WSNs and CV is used to judge whether the stable recovery in the sink node has been obtained. The detailed steps of CSCV data gathering method are shown in Fig. 2.

In CSCV, input data include measurement matrix Φ , the measurements vector, the required precision of data reconstruction τ and m the times satisfying the required precision of data reconstruction. In step 1, M and N are acquired by the dimensions of Φ and \mathbb{Y} . In step 2, collect the initial measurements and generate $\log N$ measurements \mathbb{w} as estimation set in the sink node according to the observation ways. In step 3, the measurements \mathbb{Y} are used to reconstruct the signal $\hat{\mathbf{x}}_M$ by the homotopy algorithm. In step 4, compute err by Formula (3). If $err < \tau$, goto step 7, otherwise let the counter $counter = 0$, $i = M + 1$, $y_i = 0$ and goto step 5. In step 5, a new noisy measurement y_i is acquired, then the signal $\hat{\mathbf{x}}_i$ is reconstructed by these i measurements. And the reconstruct error is estimated by (3). In step 6, if $err < \tau$, let the counter $counter = counter + 1$, $i = i + 1$, $y_i = 0$, and continue to collect $m - 1$ measurements until m times estimation errors satisfy the required precision of data reconstruction, otherwise the reconstruct errors cannot satisfy the required precision of data reconstruction, let $i = i + 1$, $y_i = 0$, and goto step 5 to collect a new noisy measurements. In step 7, output $\hat{\mathbf{x}}$ and err .

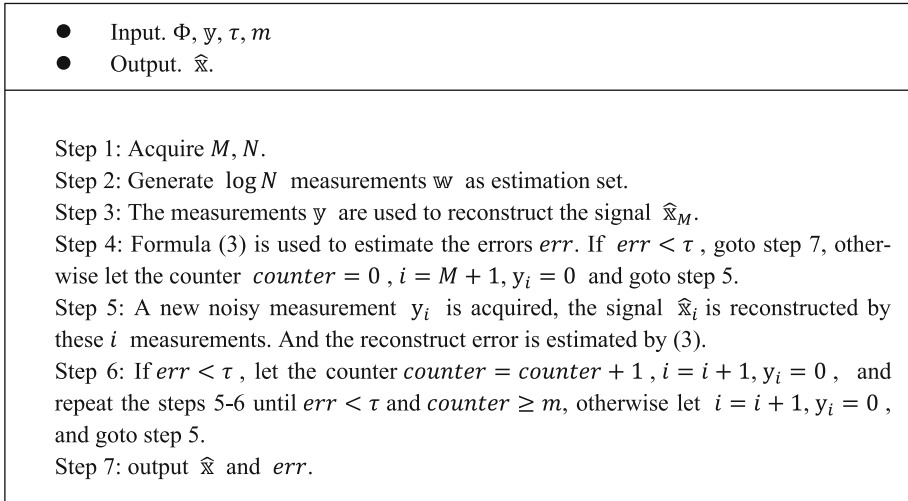


Fig. 2. The detailed steps of CSCV data gathering method.

4 Numerical Results

In the experiments, sparse signals are used as test signals due to spatial correlations of sensor data without loss of generality, and then the homotopy method is selected as the reconstruction algorithm since it is suitable to the recovery of sparse signals.

Firstly, the experiment is designed to illustrate the performance of CV estimation for CS noisy reconstruction. In this experiment, the parameters $N = 128$ and $k = 10$ are adopted, the cross validation set with $\log N$ components is adopted and the variance of the noise is 0.01. The comparison results between CV estimation error and reconstruction error are shown in Fig. 3. In this figure, the horizontal-axis represents the measurement numbers and the vertical-axis shows the reconstruction error. The solid line shows the reconstruction error of the signal, and the dashed line represents the estimation errors by CV. From this figure, we can see that CV estimation error can well approximate the reconstruction error before reaching a stable reconstruction. In this case, CV estimation error varies greatly, so it cannot be used as an end condition for obtaining stable reconstruction. Especially, the two errors are almost identical when stable reconstruction is achieved. And in this case, CV estimation error changes steadily, so it can be used as an end condition to obtain stable reconstruction. The fact concluded from the figure shows that the proposed method is correct in judging the criterion of obtaining stable reconstruction. In order to ensure the stable reconstruction results of the proposed method, the times value m satisfying the required precision of data reconstruction should be large enough. This may lead to a lot of computation run in the sink node, which is acceptable.

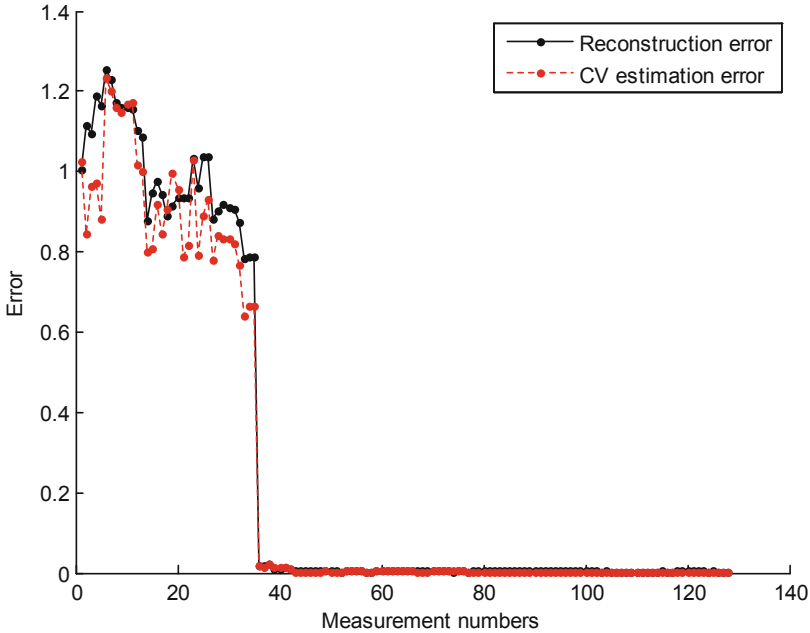


Fig. 3. Compare with CV estimation error and reconstruction error.

In order to understand the influence of the initial observation quantity on the proposed method, we designed an experiment where $N = 200$ and $k = 15$ are adopted, and the variance of the noise is 0.02. We choose uniformly initial measurement numbers M from 20 to 40 with the interval 5. The experimental results are shown in Table 1. In Table 1, the 1-st column M and the 3-rd column L represent the initial and stable measurement numbers, respectively. The 2-nd column E_1 , the 4-th column E_2 give the reconstruction errors when the initial and stable measurements are used to reconstruct sensor data, respectively. The 5-th column E_{CV} represents CV estimation error, which is very near to E_2 . When the sparsity level k is known, it is well known the fact that $3k-5k$ measurements can be taken to stably reconstruct sensor data with high probability in CS. Measurement numbers of the third column from 61–81 are consistent with the conclusion.

Table 1. Reconstruction results for different initial measurement numbers.

M	E_1	L	E_2	E_{CV}
20	0.8527	81	0.0099	0.0081
25	0.948	78	0.0087	0.0071
30	0.9433	75	0.0092	0.013
35	0.59	61	0.0085	0.0069
40	0.9967	78	0.0091	0.01

In this experiment, we will observe the influence of the signals with different sparsity level on the proposed method. In the experiment, $N = 200$ is adopted, and the variance of the noise is 0.02. We choose uniformly the signal sparsity level k from 10 to 40 with the interval 10. The experimental results are shown in Table 2. In this table, the 1-st column k is the sparsity of sensor data. The meanings of other columns are the same as Table 1. The experiment results also show that the proposed CSCV method can achieve stable reconstruction results for different sparsity of sensor data. In other words, the proposed CSCV data gathering method is insensitive to signal sparsity in the noisy data collection for WSNs.

Table 2. Reconstruction results for networks with different sparsity level.

k	M	E_1	L	E_2
10	20	0.7972	46	0.0090
20	40	0.7972	64	0.0096
30	60	0.9292	102	0.0075
40	80	0.9165	119	0.0094

Finally, we will observe the influence of the signals with different noise variance on the proposed method. The experiments with different noise variance are used to test the effectiveness of the CSCV data gathering method, and the results are shown in Table 3.

Table 3. Reconstruction results for different noise variance.

n	E_1	L	E_2
20	0.7972	46	0.0090
40	0.7972	64	0.0096
60	0.9292	102	0.0075
80	0.9165	119	0.0094

In this experiment, $N = 200$, $k = 20$ and $M = 40$. We choose uniformly the noise variance n from 0.02 to 0.08 with the interval 0.02. In this table, the 1-st column n is the noise variance, which illustrates the degree of sensor data polluted. The meanings of other columns are the same as Tables 1 and 2. The experiment results also show that the proposed CSCV method can achieve stable reconstruction results for different noise variance. In other words, the proposed CSCV data gathering method is insensitive to noise in the noisy data collection for WSNs.

In summary, the experiments above prove that the proposed method can obtain stable reconstruction results when collecting sensor data in WSNs polluted by noise, and the effectiveness of the proposed method is insensitive to the signal sparsity and noise.

5 Conclusions

In WSNs community, it is a great challenge to save and balance energy consumption of sensor nodes whose imbalance may shorten the lifetime of the network especially in noise scenarios. To efficiently gather sensor data in wireless sensor networks under noise cases, this paper proposes a CSCV data gathering method. To determine the stable recovery in sink node, CV is introduced to data gathering via CS. In the proposed method, data gathering via CS can save and balance energy consumption of sensor nodes of WSNs. And CV technique is used to judge whether these measurements can be used to stably reconstruct sensor data or not. Unlike those existing methods, CSCV data gathering method can stably reconstruct sensor data without the knowledge of the signal and/or noise. Experimental results show that the proposed method can obtain stable reconstruction results for noisy WSNs, and the effectiveness of the proposed method is insensitive to the signal sparsity and noise. CV is essentially one of statistical techniques or intelligent methods. So, future work will focus on introducing other statistical and machine learning methods into CS to collect the data under the noise case. We believe it is very promising clue by combining intelligent methods and CS in WSNs or Internet of Things fields.

Acknowledgements. This work was supported by Shanxi Province natural fund project under Grant 201801D121117, the Doctor launch scientific research projects of Datong University 2013-B-17, 2015-B-05 and ABRP of Datong under Grant 2017127.

References

1. Aguirre, E.F., Lopez-Iturri, P.S., Azpilicueta, L.T., et al.: Design and implementation of context aware applications with wireless sensor network support in urban train transportation environments. *IEEE Sens. J.* **16**(7), 169–178 (2017)
2. Akcakaya, M.F., Tarokh, V.S.: Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inf. Theory* **56**(1), 492–504 (2010)
3. Akyildiz, I.F., Su, W., Sankarasubramanian, Y., et al.: Wireless sensor networks: a survey. *Comput. Netw.* **38**(4), 393–422 (2002)
4. Baradaran, A.A.: The applications of wireless sensor networks in military environments. *Sci. J. Rev.* **4**(4), 55–70 (2015)
5. Candes, E.F., Romberg, J.S., Tao, T.T.: Near optimal signal recovery from random projections: universal encoding strategies. *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006)
6. Chen, S.F., Donoho, D.S., Saunders, M.T.: Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20**(1), 33–61 (1998)
7. Ding, X.F., Tian, Y.S., Yu, Y.T.: A real-time big data gathering algorithm based on indoor wireless sensor networks for risk analysis of industrial operations. *IEEE Trans. Ind. Inf.* **12**(3), 1232–1242 (2016)
8. Donoho, D.F.: Compressed Sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
9. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Sig. Process.* **1**(4), 586–597 (2008)

10. Khan, M.F., Pandurangan, G.S., Vullikanti, A.T.: Distributed algorithms for constructing approximate minimum spanning trees in wireless sensor networks. *IEEE Trans. Parallel Distrib. Syst.* **20**(1), 124–139 (2009)
11. Lin, H.F., Üster, H.S.: Exact and heuristic algorithms for data gathering cluster-based wireless sensor network design problem. *IEEE/ACM Trans. Netw.* **22**(3), 903–915 (2014)
12. Lindsey, S.F., Raghavendra, C.S., Sivalingam, K.M.T.: Data gathering algorithms in sensor networks using energy metrics. *IEEE Trans. Parallel Distrib. Syst.* **13**(9), 924–935 (2002)
13. Mallet, S.F., Zhang, Z.S.: Matching pursuits with time-frequency dictionaries. *IEEE Trans. Sig. Process.* **41**(12), 3397–3415 (1993)
14. Song, X., Li, Y.: Data gathering in wireless sensor networks via regular low density parity check matrix. *IEEE/CAA J. Autom. Sin.* **5**(1), 83–91 (2018)
15. Tibshirani, R.F.: Regression shrinkage and selection via the lasso. *J. R. Stat.* **58**(1), 267–288 (1996)
16. Xiao, Y.F., Yang, J.S.: A fast algorithm for total variation image reconstruction from random projections. *Inverse Prob. Imaging* **6**(3), 547–563 (2017)
17. Younis, O.F., Fahmy, S.S.: HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans. Mobile Comput.* **3**(4), 366–379 (2004)
18. Zhang, J., Chen, L., Boufounosl, P.T.: On the theoretical analysis of cross validation in compressive sensing. In: 2014 Conference, ICASSP, pp. 3370–3374. IEEE, Florence (2014)
19. Zhang, P.F., Wang, S.S., Guol, K.T.: A secure data collection scheme based on compressive sensing in wireless sensor networks. *Ad Hoc Netw.* **70**(1), 73–84 (2018)
20. Zhu, B., Suzuki, J., Boonma, P.: Evolutionary and noise-aware data gathering for wireless sensor networks. In: Suzuki, J., Nakano, T. (eds.) *BIONETICS 2010*. LNICST, vol. 87, pp. 32–39. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32615-8_5



Fuzzy-K: Energy Efficient Fuzzy Clustering Routing Protocol Based on Cross-Technology Communication in Wireless Sensor Network

Yue Yu¹, Fanrong Meng^{1,2(✉)}, and Ming Li^{1,2}

¹ School of Computer Science and Technology,
China University of Mining and Technology, Xuzhou 221116, China
mengfr@cumt.edu.cn

² Mine Digitization Engineering Research Center of Ministry of Education,
Xuzhou, People's Republic of China

Abstract. The rapid development of wireless sensor networks (WSN) in various fields has brought great convenience. Since most wireless sensor nodes are powered by batteries, energy efficiency is very important for WSN, and many existing routing protocols aim to reduce energy consumption. At present, the emergence of cross technology communication (CTC) enables direct communication between heterogeneous nodes at the physical layer. Therefore, new routing algorithms need to be designed for WSN based on CTC, which considers the heterogeneous characteristics of sensor nodes such as the energy heterogeneity, the mobility of LTE nodes, etc. In this paper, we propose an energy efficient fuzzy clustering routing protocol based on CTC, which is named as Fuzzy-K. Different from other protocols, Fuzzy-K first uses k-means algorithm to form balanced clusters and then select CH (cluster head). In the proposed protocol, the Mamdani fuzzy inference system (FIS) is used twice to select the initial cluster center and the final CH. The input parameters of these two systems are obviously different, which considers the differences in frame length, mobility and other heterogeneous characters between nodes. Simulation results of three different network topologies show that compared with LEACH, EEHCCP, TEAR and DUCF, Fuzzy-K protocol has better performance in extending network life cycle, balancing network load and improving network throughput. And the average value of the rounds when first node dies could be 27% higher than the protocols mentioned above. What's more, the proposed protocol is scalable across a range of situation by changing parameters of the FIS.

Keywords: CTC · Wireless Sensor Network (WSN) · Clustering · Fuzzy inference system · K-means algorithm

1 Introduction

Wireless Sensor Network (WSN) is a type of wireless network composed of a large number of static or mobile sensors distributed in a self-organizing and multi-hop way, which collaboratively senses, collects, processes and transmits the information of the

perceived object. There are many types of wireless sensors that can detect various phenomena of the surrounding environment, including earthquake, temperature, humidity, noise, light intensity, pressure, soil composition, size, speed and direction of moving objects [1]. From this perspective, WSN has a wide range of potential applications in earthquake [2], temperature [3], humidity [4], noise [5], light [6], pressure and medical care [7, 8] and other fields. Each wireless sensor node is composed of four main parts: sensor unit, processing unit, energy supply unit and transceiver unit [9]. Since the sensor nodes are generally powered by battery, the improvement of energy efficiency has always been a very important topic for the application of WSN. As for the network level, it is of great significance to design an efficient routing protocol to maximize the entire lifetime of WSN.

On the other hand, with the progress of wireless communication technology, various fields of WSN are developing rapidly. Traditional wireless bridging technologies are implemented indirectly through multi-radio gateways, which brings additional hardware costs, deployment complexity, and doubles the traffic into and out of these gateways [10]. Fortunately, many recent studies have shown that great progress has been made in CTC, which supports direct communication between different wireless signals [11–13]. By processing the frame structure of wireless signals on the physical layer, CTC enables different types of wireless devices to exchange data with each other. As a result, such transparent mapping technology has eliminated the issues addressed above.

In terms of the upper layer of the network, most of the current routing protocols are designed for the homogeneous network. The differences of node energy and the change of position are not considered. Therefore, they cannot be well adapted to the difference of data packets in CTC network. Based on the increasingly mature CTC technology, a new routing protocol named Fuzzy-K is proposed in this paper, which commits to maximize energy efficiency and extend the lifetime of the network. This protocol integrates the fuzzy inference system (FIS) and the classical k-means algorithm in the clustering field. In the process of clustering, the position and distribution of the nodes are set as the input of the FIS to accommodate the mobility of LTE nodes. Besides, the difference in initial energy is fully considered. During communication among different nodes, the differences of transmission power and frame length are considered in the energy model. Other normal parameters such as the distance to the base station (BS), the number of neighbor nodes, the selection history of nodes, etc. are considered to select the cluster head (CH) reasonably.

The contributions of this paper can be summarized as follows:

- A new energy efficient routing protocol is designed for heterogeneous WSN based on CTC, effectively extending the network life cycle and improves the network throughput.
- A routing scheme is proposed for WSNs, in which, the order of the process of selecting CH and forming clusters is inverted.
- Mamdani fuzzy inference system is used twice to choose the appropriate initial clustering centers and final CHs. Finally, through the simulation of three different network topologies, it is proved that the proposed protocol performs well in extending the network life cycle and balancing clustering.

The rest of this paper is organized as follows. Section 2 discusses several representative CTC technologies and typical clustering routing protocols. Section 3 describes the heterogeneous network model and energy consumption model. The proposed protocol is introduced in Sect. 4. Section 5 explains the simulation setup. The simulation results are analyzed in Sect. 6. Finally, the conclusion and some suggestions for the future work are presented in Sect. 7.

2 Related Works

2.1 Representative Technologies in CTC

CTC [10, 14–16] technology, which modulates signals on the physical layer of the network so that different wireless signals can communicate directly, has made great progress. WEBee [10] (for WiFi emulated ZigBee) realizes high-throughput CTC via physical-level emulation. It chooses the payload of the WiFi frame in such a way that the ZigBee device transparently identifies a portion of the WiFi frame as a legitimate ZigBee frame. Since the WiFi frame structure and the WiFi signal transmitter hardware are not modified, the WiFi receiver can still receive the normal WiFi frame legally. Thus, the direct communication is realized between WiFi devices and ZigBee devices. The experimental results show that in noisy environment, the reliability rate of WEBee can reach above 99%, and the fastest speed can reach 126 Kbps.

LtFi [14] is the first system to allow CTC between LTE-U and WiFi. Piotr et al. proposed a two-step approach: an innovative side channel is used on its air-interface LTE-U BS to broadcast the connection and identification information to adjacent WiFi nodes. Then, the node is used to create a two-way control channel on the wired return journey. The channel based on the air interface can reach up to 665 bps. The simple LtFi is fully compatible with LTE-U and COTS WiFi hardware at the same time. According to [14], the LtFi system can provide reliable data transmission even in the multi-interference wireless environment where receiving power level for LTE-U drops down to -92 dBm.

2.2 Routing Protocols

Routing protocols can be classified into plane and hierarchical routing protocols. In this paper, the hierarchical routing protocols are discussed. Classic routing protocols such as LEACH [17], LEACH-C [18], LEACH-DT [19], protocols based on fuzzy system such as DUCF [20], Type2FL [21], SIF [22], ALSPR [23], FSFLA [24], as well as heterogeneous network protocols such as TEAR [25], EEHCCP [26], etc., can reduce the network energy consumption and improve throughput in WSN.

LEACH [17] (Low Energy Adaptive Clustering Hierarchy) is the first hierarchical routing protocol, which is a layered, probabilistic, distributed and single-hop protocol based on random distribution of sensor nodes. The whole WSN is divided into different clusters, and each cluster consists of a CH node and several cluster member nodes. In order to manage cluster member nodes, the CH node is also responsible for processing the information sent by cluster member nodes and transmitting it to the BS. The life

cycle of the network is divided into several rounds, each of which is divided into two steps: setup step and steady-state step. At the beginning of each round, the cluster node is selected in the setup step, and then other nodes form clusters on the principle of proximity. Then it enters the steady-state step, member nodes in the cluster transmit the data packets to the CH node, which will process them and send them to the BS. LEACH protocol achieves the purpose of balancing network load and reducing network energy consumption through the periodic CHs replacement process, which effectively extends the life cycle of the entire network.

DUCF (Distributed load balancing Unequal Clustering in wireless sensor networks using Fuzzy approach, DUCF) [20] is a distributed unequal clustering algorithm based on fuzzy logic. The FIS used in DUCF takes the residual energy, node degree and distance to the BS as the input variables for CH selection. There are two output parameters – Chance and Size. At the beginning of each round, each node is calculated, and the node with the largest Chance parameter is selected as the CH. At the same time, non-CH nodes are joined to the clusters until the number of members in the cluster reaches the Size threshold. Eventually, the remaining nodes automatically become new CHs. The simulation results show that this algorithm forms unequal clusters, which ensures the load balance between clusters, and further improves the network life cycle.

Type2FL (Energy Efficient Clustering Algorithm for Multi-Hop Wireless Sensor Network Using Type-2 Fuzzy Logic) [21] propose a clustering algorithm on the basis of interval type-2 fuzzy logic model, which handles uncertain level decision better than T1FL model. The T2FL model is used to improve the routing algorithm by selecting a CH efficiently. The whole sensor network is divided into numbers of levels. At each level, CH is selected based on T2FL Model. For each node, residual energy, distance to BS and concentration have been considered. Each CH sends the data to the next level which finally reaches the BS.

TEAR (Traffic and Energy Aware Routing for Heterogeneous Wireless Sensor Networks) [25] considers the heterogeneity of nodes to achieve the optimal utilization of resources. This protocol considers sensor nodes with difference in random initial energy and random data generation rate (flow rate), establishing a real clustering-based WSN model suitable for heterogeneous sensing applications. Based on the classical LEACH protocol, the algorithm improves the process of CH selection. After a series of formula derivation, a new threshold calculation method is proposed.

EEHCCP (Efficient Energy Heterogeneous Circular-field Clustering Protocol) [26] is a heterogeneous circular-field routing algorithm with high energy efficient. The protocol deploys two layers of heterogeneous energy nodes in different areas: normal nodes and advanced nodes. The advanced node has higher energy and communicates directly with the BS, and normal nodes communicate according to the classical LEACH protocol. The simulation results show that this protocol improves the network lifetime and throughput.

Nodes in all the above protocols merely choose the nearest CH to form clusters, so that the clusters are not considered as a whole entirety. Besides, whether the node is chosen as the CH or not and whether the common node is clustered depends on the nature of the node itself, ignoring the aspect of minimizing the distance within the cluster. On the other hand, with the introduction of CTC technology, different types of nodes whose frame lengths are slightly different coexist in the whole network. The

heterogeneity of nodes makes the energy consumption during the process of information transmission quite different. Furthermore, LTE nodes are mobile so that their positions often change. All the above factors make the existing protocols not well applicable to the WSN based on CTC. Hence, in this paper, a new energy efficient routing protocol Fuzzy-K is proposed, which is designed for heterogeneous networks based on CTC to maximize network life cycle, balance network load and improve network throughput.

3 System Model

3.1 Network Model

The protocol proposed in this paper is based on the recent significant progress in the field of CTC, such as WEBee [10], LtFi [14] and other new physical layer technology. The heterogeneity in WSN here is the difference between real signals in the physical layer instead of simple difference in energy and traffic, which should be considered in the routing protocol. We consider a single-hop cluster-based wireless sensor heterogeneous network [22], and the network model is shown in Fig. 1.

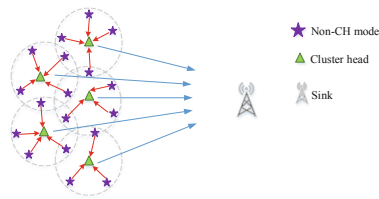


Fig. 1. Single-hop clustering-based WSN model.

According to CTC, it is assumed that all nodes in the network can communicate with each other, and the data is finally gathered in the BS. All LTE nodes are equipped with GPS (global positioning system), which can be used to locate its own position and the position of the BS after location updating in each round. All nodes can adjust the transmission power according to the distance from the receiving nodes. Each node should hibernate after transmission and wake up in the next round. All nodes die simply because of energy depletion.

3.2 Energy Model

The network model proposed in this paper is a physically heterogeneous network, in which there are three types of nodes: Zigbee, WiFi and LTE. As for the energy consumption model, the same model as the LEACH protocol is used to calculate the energy consumption [17]. The energy consumed in transmitting an m_1 -bits message on distance d is given by Eq. (1):

$$ETx(mi, d) = \begin{cases} m_i \cdot \beta \cdot E_{ele} + m_i \cdot \varepsilon_{fs} \cdot d^2 & \text{if } d < d_0 \\ m_i \cdot \beta \cdot E_{ele} + m_i \cdot \varepsilon_{mp} \cdot d^4 & \text{if } d \geq d_0 \end{cases} \quad (1)$$

where E_{ele} (nJ/bit) is the energy consumed in electronic circuit by the receiver or the sender when sending or receiving per bit, ε_{fs} and ε_{mp} are fixed radio parameters in free space and multipath fading channel used for the transmitter amplifier, and d_0 is the distance threshold defined by Eq. (2):

$$d_0 = \sqrt{\frac{\varepsilon_{fs}}{\varepsilon_{mp}}} \quad (2)$$

β is the energy consumed parameter of the wireless sensor nodes after normalization. The energy consumption during transmission of the three kind of nodes is different [27–30]. According to the throughput and transmitted power, the relation of three different nodes in energy consumption during network operation could be calculated. The network throughput per unit energy consumption [31] is defined by Eq. (3):

$$EE = \frac{\log_2(1 + p \cdot h)}{a_{site} \cdot P + b_{site}} \quad (3)$$

where p is the transmission power, h is the channel-to-interference-plus-noise ratio; $a_{site} \cdot P$ represents the transmission power consumption; b_{site} is the circuit power consumption; a_{site} is the power-conversion efficiency, accounting for the power amplifier efficiency, feeder loss, extra loss in transmission-related cooling, etc.

The energy consumed in receiving an m_i -bits packet is given by Eq. (4):

$$E_{Rx}(m_i) = m_i \cdot \beta \cdot E_{ele} \quad (4)$$

4 Proposed Protocol

The proposed protocol named Fuzzy-K balances network load, reduces energy consumption and improves energy efficiency by reasonably clustering and selecting the optimal CH node, so as to maximize the life cycle of the whole network.

4.1 Overall Description of Fuzzy-K

The Fuzzy-K protocol proposed in this paper is developed based on the core idea of the protocol LEACH. However, k-means algorithm [32, 33] is used for clustering before selecting CHs different from other protocols. The initial clustering center is determined by FIS-I. After k-means algorithm, the total sensor nodes are reasonably divided into several clusters. The characteristics of clusters are that the distances within clusters are minimized and the distances between them are maximized. Finally, the final CHs are selected within clusters based on the second FIS with other different inputs. The overall

operation and time allocation of each round are shown in Fig. 2, and the overall flow chart is shown in Fig. 3.

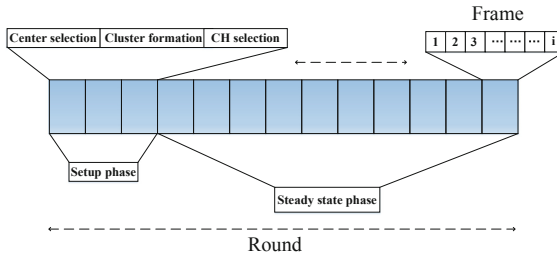


Fig. 2. Operation and time allocation of each round.

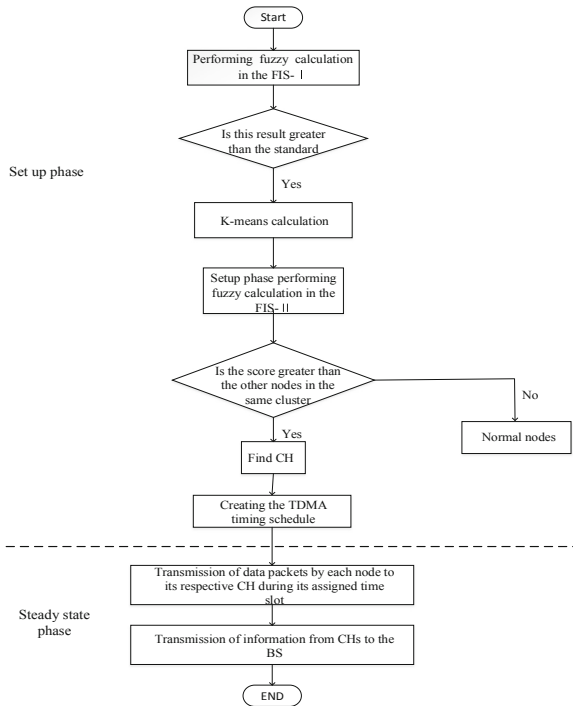


Fig. 3. Flowchart in operating steps.

4.2 FIS

Two FIS models are adopted: FIS-I and FIS-II respectively. The FIS-I is designed for all nodes in the network to produce the initial center for subsequent operation of clustering; FIS-II is applied after clustering to select the final CH nodes.

FIS-I. (1) Input A: Residual energy represents the residual energy of node i in the heterogeneous network. (2) Input B: Neighbor represents the number of neighbor nodes around node i . (3) Input C: Distance to BS represents the distance between node i and the BS. (4) Output: Put the three kinds of data above into the first FIS, the first output named Output-I will be obtained, which represents the ability of nodes in CH selection. We select the nodes whose output is not less than a certain standard as the initial center for clustering and the initial input of k-means clustering algorithm. Moreover, the standard is artificial regulated so that we can make appropriate adjustment in order to get better results according to different application scenarios. And the Output-I is divided into 9 levels, that are very high, high, rather high, high, medium, medium, low medium, rather low, low, very low as shown in Fig. 4.

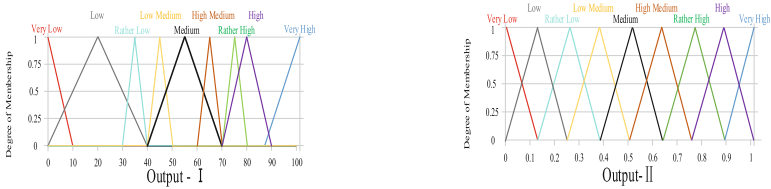


Fig. 4. Membership function for output variable ‘Output-I’ and ‘Output-II’.

FIS-II. After the operation of FIS-I, the nodes in heterogeneous network are reasonably divided into several clusters. Then the final CH nodes are determined by FIS-II. The parameters are different from the first one. (1) Input D: Residual Energy represents the residual energy of node i . (2) Input E: Distance to all nodes represents the sum of distances between node i and other nodes in the cluster. (3) Input F: History represents the “history” of the nodes in the heterogeneous network, that is, the number that the node has been selected as the CH. (4) Output: Put the three kinds of data above into the FIS-II, an output named Output-II is given, which is used as the selection rules for the final CH. The nodes will be arranged according to the value of Output-II. Nodes with the highest value is to be selected as the final CH. In this paper, the Output -II is divided into nine levels as shown in Fig. 4.

5 Simulation Setup

Three scenarios are considered in this paper to evaluate the performance of the proposed protocol. The simple topologies are shown in Figs. 5, 6 and 7.

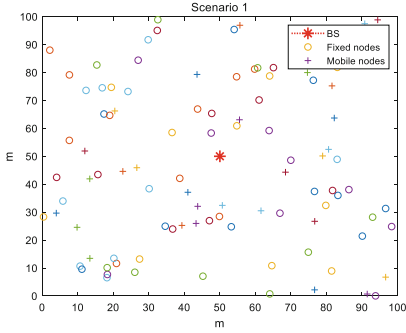


Fig. 5. Scenario 1: BS at (50, 50).

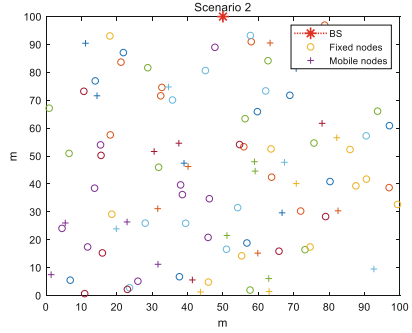


Fig. 6. Scenario 2: BS at (50, 100).

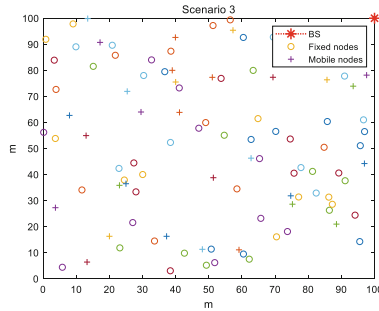


Fig. 7. Scenario 3: BS at (100, 100).

The heterogeneous network is composed of three types of nodes: WiFi, Zigbee and LTE. According to the actual heterogeneous network, the ratio of the nodes is set as 4:3:3. Among them, since the characters are quite different according to [27–30], the initial energy of WiFi nodes, Zigbee nodes and LTE nodes is set as 2:1:3. According to the throughput and transmitted power, the energy consumed parameter of WiFi nodes, Zigbee nodes and LTE nodes is set as 0.19, 0.12 and 0.69 respectively. In order to avoid the inconvenience brought by the actual value to the simulation, the parameters are normalized. Other network parameter settings are shown in Table 1.

Table 1. Simulation parameters and values.

Parameter	Value
WSN area($R \times R$)	100 m \times 100 m
Number of sensor nodes(N)	100
Initial energy of Zigbee nodes (E_0)	0.5 J
Energy consumed in Tx/Rx electronics (E_{ele})	50 nJ/bit
Energy consumed parameter of nodes ($\beta_W/\beta_Z/\beta_L$)	0.19/0.12/0.69
Tx Amplifier energy dissipation in free space scenario (ϵ_{fs})	100 pJ/bit/m ²
Tx Amplifier energy dissipation in Multipath scenario (ϵ_{mp})	0.013 pJ/bit/m ⁴
Energy in Data Aggregation (E_{DA})	5 nJ/bit/signal
Data packet size	5000 bit
Control packet size	50 bit

6 Simulation Results and Analysis

The proposed protocol Fuzzy-K will be compared with LEACH [17], DUCF [20], TEAR [25], and EEHCCP [26]. In each scenario, the simulation is carried out for 5 times and the results are averaged to be compared between different algorithms. All the protocols mentioned above are evaluated according to four factors: number of alive nodes, number of received packets, mean maximum intra-cluster distance and mean intra-cluster distance.

6.1 The Number of Alive Nodes in Each Round

For a WSN with limited energy, the life cycle of the whole network and the performance of the routing protocol applied can be well reflected by the number of the alive nodes in each round. The number of alive nodes in all rounds are qualified in Fig. 8. Fuzzy-K protocol performs better than the other four algorithms. This is because clustering and choosing the CHs are in opposite order compared to other protocols. In the process of clustering, the position and distribution of nodes are fully considered. Balanced clusters can be formed according to the location of nodes in different scenarios. Therefore, the energy consumed by intra-cluster communication will be greatly reduced.

The number of rounds using the proposed protocol is less than some other protocols such as EEHCCP and TEAR sometime while the number of active nodes is reduced to 20 or less. However, the situation of communication in the network will be slightly affected. Even for some applications in special scenarios, the network with few alive nodes has no use at that time.

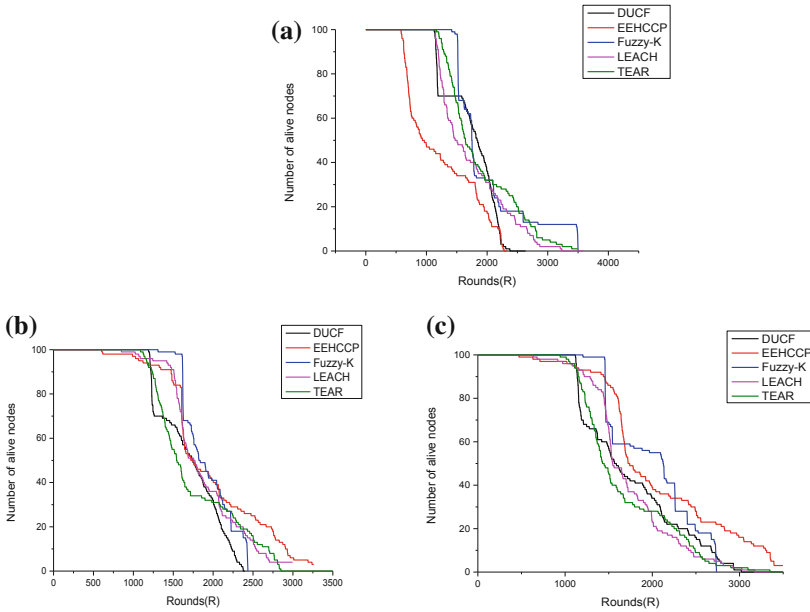


Fig. 8. (a) Number of alive nodes in all rounds in scenario 1. (b) Number of alive nodes in all rounds in scenario 2. (c) Number of alive nodes in all rounds in scenario 3.

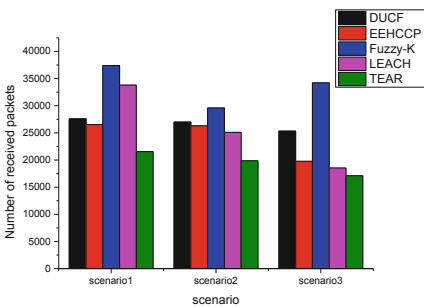


Fig. 9. Number of received packets in different scenarios.

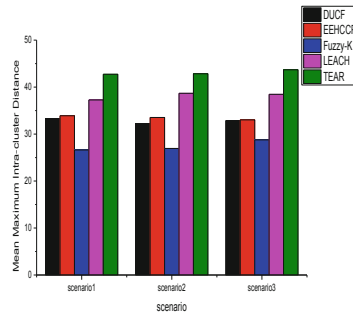


Fig. 10. Mean maximum intra-cluster distance in different scenarios

6.2 Total Number of Data Packets Received by the BS

The number of data packets received by the BS before the FND (First Node Dies), that is, the round number after the first node death is presented in Fig. 9 and Table 2. To some extent, the communication capability of the whole network can be measured by the total number of packets received by the BS. For different protocols, the number of rounds may be similar, but the number of packets received by the BS before FND may

vary. By comparing of the results in three topologies, the number of packets received by Fuzzy-K protocol in the same scenario is higher than that of other protocols, which indicates that this protocol can effectively improve the network throughput. In addition, the throughput of the Fuzzy-K protocol is the highest under the different network topologies. That is, the proposed protocol performs well in different situations. It can be summarized that the Fuzzy-K protocol can better adapt to different network scenarios.

Table 2. Number of received packets in different scenarios

Protocol	Scenario 1	Scenario 2	Scenario 3
DUCF	27578.6	26987.1	25315.8
EEHCCP	26530.3	26297.4	19764.5
Fuzzy-K	37407.4	29587.5	34204.7
LEACH	33799.1	25097.1	18533.3
TEAR	21555.1	19865.7	17109.5

6.3 The Mean Maximum Intra-cluster Distance in All Rounds

The mean maximum intra-cluster distance for all protocols is presented in Fig. 10. In the proposed protocol, k-means algorithm is used to form clusters, instead of choosing the nearest CH for non-CH nodes. The number and size of clusters are adjusted adaptively according to the distribution of nodes in the whole network so that the intra-cluster distances can be effectively reduced. As for clustering protocols, the average of the intra-cluster distance is an important parameter to measure the clustering results and the selection of the CH. The reduction of average intra-cluster distance can save the energy consumed by intra-cluster communication so that the energy efficiency of the whole network could be improved. As is shown, the mean maximum intra-cluster distance in the network applied in the proposed protocol is lowest and has little change in different network topologies, which is also an important factor on the life cycle of the whole network.

6.4 The Mean Intra-cluster Distance in All Rounds

Figure 11 reflect the mean intra-cluster distance in the three scenarios. This parameter refers to the average value of the maximum distance between nodes in the same cluster. As it can be seen from the box graph, the distribution of average distance within the cluster of Fuzzy-K protocol is more concentrated and lower overall. It should be noted that the network model in this paper is special that there is a certain proportion of LTE nodes which are mobile. By applying k-means clustering algorithm, the moving nodes will be clustered according to their positions during the initialization of each round of network. After clustering, the appropriate CH will be re-selected. This makes the clustering more balanced and the selection of CH in each cluster more reasonable so that the load of network is more balanced.

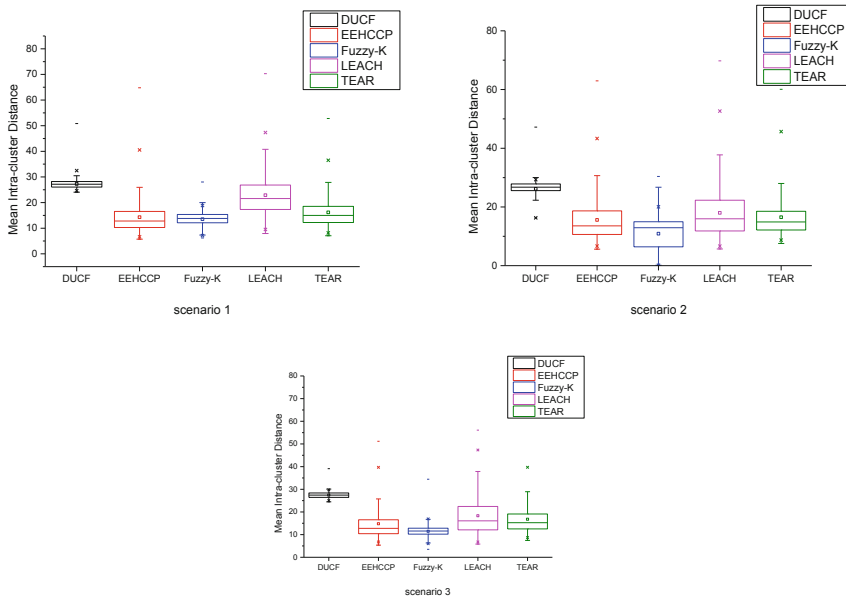


Fig. 11. Mean intra-cluster distance for all rounds in different scenarios.

7 Conclusion

With the rapid development of WSN, it is still very important to improve the energy efficiency, maximize the life cycle and balance the load of the whole network. Aiming at the physical heterogeneous network based on CTC, the Fuzzy-K protocol is designed according to the differences caused by the coexistence of multiple wireless nodes. In this proposed protocol, the residual energy of nodes, the distance from the BS, mobility, history of nodes and other factors are considered. Besides, the k-means clustering algorithm is introduced to balance clustering. And the FIS is used to select the CH nodes and properly avoid the limitations of the k-means algorithm. Through exchanging the order of clustering and CH selecting, the intra-cluster distance can be reduced efficiently. FIS is used twice to make the selection of initial centers and CHs more reasonable. Compared with the classical LEACH protocol, the latest heterogeneous network protocols EEHCCP and TEAR, and DUCF protocol based on fuzzy theory, the Fuzzy-K protocol performs well in the life cycle, throughput and intra-cluster distance of the network.

However, there are also some limitations. The proposed protocol is designed based on the idea of single-hop hierarchical routing algorithm, and multi-hop paths are not taken into consideration. In the future, we will continue to improve the performance of Fuzzy-K protocol by designing multi-path and multi-hop algorithm to adapt to a larger scale network topology.

Funding. This research was funded by The National Key Research and Development Program of China, grant number 2016YFC0600908 and The National Natural Science Foundation of China, grant number 51874302.

References

1. Ye, W.: Research on the application of Internet of Things technology in intelligent home. In: 5th International Conference on Mechatronics, Materials, Chemistry and Computer Engineering (ICMMCE) Location: Chongqing, Peoples Republic of China, 24–25 July 2017
2. Alphonsa, A., Ravi, G.: Earthquake early warning system by IOT using Wireless sensor networks. In: 2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1201–1205. IEEE (2016)
3. Badia-Melis, R., Garcia-Hierro, J., Ruiz-Garcia, L., Jiménez-Ariza, T., Villalba, J.I.R., Barreiro, P.: Assessing the dynamic behavior of WSN motes and RFID semi-passive tags for temperature monitoring. *Comput. Electron. Agric.* **103**, 11–16 (2014)
4. Udaykumar, R.Y.: Development of WSN system for precision agriculture. In: 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–5. IEEE (2015)
5. Kivelä, I., Hakala, I.: Area-based environmental noise measurements with a wireless sensor network. In: Proceedings of the Euronoise, pp. 218–220 (2015)
6. Lavric, A., Popa, V., Sfichi, S.: Street lighting control system based on large-scale WSN: a step towards a smart city. In: 2014 International Conference and Exposition on Electrical and Power Engineering (EPE), pp. 673–676. IEEE (2014)
7. Saeed, H., Ali, S., Rashid, S., Qaisar, S., Felemban, E.: Reliable monitoring of oil and gas pipelines using wireless sensor network (WSN)—REMONG. In: 2014 9th International Conference on System of Systems Engineering (SOSE), pp. 230–235. IEEE (2014)
8. Liang, T., Yuan, Y.J.: Wearable medical monitoring systems based on wireless networks: a review. *IEEE Sens. J.* **16**(23), 8186–8199 (2016)
9. Thilagavathi, S., GeethaPriya, C.: Study on wireless sensor networks - a comprehensive approach. In: 7th IEEE International Conference on Communication and Signal Processing (IEEE ICCSP) Adhiparasakthi Engineering College, Melmaruvathur, India 03–05 April 2018
10. Li, Z., He, T.: WEBee: physical-layer cross-technology communication via emulation. In: 23rd Annual International Conference on Mobile Computing and Networking (MobiCom) Location: Snowbird, UT, 16–20 October 2017
11. Yin, Z., Jiang, W., Kim, S.M., He, T.: C-morse: cross-technology communication with transparent morse coding. In: INFOCOM (2017)
12. Jiang, W., Yin, Z., Kim, S.M., He, T.: Transparent cross-technology communication over data traffic. In: INFOCOM (2017)
13. Kim, S.M., He, T.: FreeBee: cross-technology communication via free side-channel. In: MobiCom 2015, pp. 317–330. ACM, New York. <https://doi.org/10.1145/2789168.2790098>
14. Gawłowicz, P., Zubow, A., Wolisz, A.: Enabling cross-technology communication between LTE unlicensed and WiFi. In: IEEE Conference on Computer Communications (IEEE INFOCOM) Location: Honolulu, HI, 15–19 April 2018
15. Wei, W., He, S., Sun, L.: Cross-technology communications for heterogeneous IoT devices through artificial doppler shifts. *IEEE Trans. Wirel. Commun.* **18**, 796–806 (2019)

16. Demin, G., Shuo, Z., Fuquan, Z.: RowBee: a routing protocol based on cross-technology communication for energy-harvesting wireless sensor networks. *IEEE Access* **7**, 40663–40673 (2019)
17. Singh, K.: WSN LEACH based protocols: a structural analysis. In: 2015 International Conference and Workshop on Computing and Communication (IEMCON), Vancouver, CANADA, 15–17 October 2015
18. Ge, Y., Jie, K., Kun, T.: The improved LEACH-C protocol with the cuckoo search algorithm. In: International Academic Conference on Computer Networks and Communication Technology (CNCT), Xiamen, Peoples Republic of China, 16–18 December 2016
19. Baranidharan, B., Santhi, B.: DUCF: distributed load balancing unequal clustering in wireless sensor networks using fuzzy approach. *Appl. Soft Comput.* **40**, 495–506 (2016)
20. Darabkh, K.A., Zomot, J.N.: An improved cluster head selection algorithm for wireless sensor networks. In: 14th IEEE International Wireless Communications and Mobile Computing Conference (IWCMC), Limassol, CYPRUS, 25–29 June 2018
21. Nayak, P., Vathasavai, B.: Energy efficient clustering algorithm for multi-hop wireless sensor network using type-2 fuzzy logic. *IEEE Sens. J.* **17**(14), 4492–4499 (2017)
22. Zahedi, Z.M., Akbari, R., Shokouhifar, M., Safaei, F., Jalali, A.: Swarm intelligence based fuzzy routing protocol for clustered wireless sensor networks. *Expert Syst. Appl.* **55**, 313–328 (2016)
23. Shokouhifar, M., Jalali, A.: A new evolutionary based application specific routing protocol for clustered wireless sensor networks. *Int. J. Electr. Commun. (AEÜ)* **69**, 432–441 (2015)
24. Fakhrosadat, F., Rafsanjani, M.K.: Memetic fuzzy clustering protocol for wireless sensor networks: shuffled frog leaping algorithm. *Appl. Soft Comput.* **71**, 568–590 (2018)
25. Sharma, D., Bhondekar, A.P.: Traffic and energy aware routing for heterogeneous wireless sensor networks. *IEEE Commun. Lett.* **22**, 1608–1611 (2018)
26. Chithra, A., Shantha Selva Kumari, R.: A new energy efficient clustering protocol for a novel concentric circular wireless sensor network. *Wirel. Personal Commun.* **103**, 2455–2473 (2018)
27. Denkovski, D., Rakovic, V., Atanasovski, V.: Power and channel optimization for WiFi networks based on REM data. *Wirel. Personal Commun.* **97**, 1753–1779 (2017)
28. Pan, G., He, J., Wu, Q.: Automatic stabilization of zigbee network. In: International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, Peoples Republic of China, 26–28 May (2018)
29. Chen, S., Hu, J., Shi, Y.: LTE-V: a TD-LTE-based V2X solution for future vehicular network. *IEEE IoT J.* **3**, 997–1005 (2016)
30. Yunas, S.F., Valkama, M., Niemelä, J.: Spectral and energy efficiency of ultra-dense networks under different deployment strategies. *IEEE Commun. Mag.* **53**, 90–100 (2015)
31. Ming, L., Pengpeng, C., Shouwan, G.: Cooperative game-based energy efficiency management over ultra-dense wireless cellular networks, pp. 1424–8220. SEP, Sensors (2016)
32. Wang, X., Bai, Y.: The global Minmax k-means algorithm. *Springerplus* **5**, 1665 (2016)
33. Zhang, H., Yu, H., Li, Y.: Improved K-means algorithm based on the clustering reliability analysis. In: International Symposium on Computers and Informatics (ISCI), Beijing, Peoples Republic of China, 17–18 January 2015



An Improved Method of Pending Interest Table in Named Data Networking

Peiyuan Gu^{1,2}, Yabin Xu^{1,2(✉)}, and Tian Song³

¹ Beijing Key Laboratory of Network Culture and Digital Communication, Beijing Information Science and Technology University, Beijing 100101, China
peiyuangu@mail.bistu.edu.cn, xyb@bistu.edu.cn

² School of Computer, Beijing Information Science and Technology University, Beijing 100101, China

³ School of Computer, Beijing Institute of Technology, Beijing 100081, China
songtian@bit.edu.cn

Abstract. In Named Data Networking (NDN), Pending Interest Table (PIT) is proposed to record the forwarding information of interest packets forwarded but not responded. Each incoming interest packet or data packet needs to be queried and processed in PIT, and the overhead would rise as the scale of PIT increases. Therefore, PIT is required to have a very high processing speed. To effectively improve the forwarding efficiency of PIT in NDN, a new architecture of the PIT using a hot table to achieve prefix grading is designed and implemented. The concept of “prefix value” is proposed to determine the value of a prefix carrying the information content of an interest packet, and to store and prioritize the prefix information with a higher value. The results of the comparison experiment show that the architecture of the PIT with a hot table can significantly improve processing speed of the PIT and accelerate forwarding efficiency of the NDN node.

Keywords: Named Data Network · Pending Interest Table · Prefix value · Hot table

1 Introduction

The Named Data Network (NDN) is a typical implementation of Information Centric Networking (ICN) [1]. Unlike traditional IP networks, NDN focuses on the content itself rather than the location of the content [2]. It is routed based on the content name without relying on IP-like location information; the NDN content caching mechanism also greatly reduces the network load of the node, making it more suitable for the ubiquitous Internet content sharing mode of the data source [3].

Since the NDN relies on the Pending Interest Table (PIT) to achieve efficient forwarding of data, and each incoming interest packet or data packet needs to be queried and processed in the PIT [4], to a certain extent, whether the design of the PIT is scientific and reasonable can directly determine the efficiency of data forwarding.

To achieve a quick match of PIT, researchers have proposed some effective methods, which are mainly divided into two categories.

The first category is to improve the performance of PIT by optimizing PIT search, replace, delete and other strategies. Paper [5] first proposed the concept of Name Component Encoding (NCE), which is used to reduce the size of PIT and to meet the requirements of access frequency. This mechanism can reduce the number of component encoding and the encoding length of each component, but does not reduce the correctness of the longest name prefix match. The name component encoding mechanism separates the encoding process from the longest prefix match, making it possible to use parallel processing techniques to speed up name lookups. Also, this method limits the name lookup time to the upper limit of the maximum time between the component encoding process and the longest encoding name prefix match. Based on this method, the paper [6] put forward the idea of Name Prefix Trie (NPT), which further improves the storage and operation performance of PIT. However, this method requires time to encode and construct the matching tree, so it is more suitable for large-scale NDN networks. For small-scale NDN networks, search performance may not be significantly increased due to additional coding time.

In [7], an instant-triggered PIT entry aging method is proposed, which can achieve an extremely high-precision PIT entry timeout judgment with only a small amount of storage space. In [8], the one-to-one data interaction mode of interest packets and data packets in the named data network are changed, and a method of returning multiple data packets by sending one interest packet is realized, thereby improving the efficiency of routing and forwarding. Paper [9] dynamically changes the timeout period for each interest packet entering the PIT, preventing the phenomenon that the unresponsive interest packet occupies excess space in the PIT. Paper [10] changes the replacement method when the item overflows in the PIT so that the entry that is least likely to be responded is replaced with the highest priority.

Although this kind of method implemented strategy optimization on a certain aspect of PIT, limitations still exist. It does not improve the performance of the PIT at the root cause, and cannot fully adapt to all NDN network environments.

The second category is to optimize PIT performance by changing the PIT architecture. Paper [11] uses Bloom Filter (BF) for space compression of PIT to improve the lookup performance. Paper [12] and [13] proposed the concepts of MaPIT and DiPIT on the basis of Bloom Filter. The former effectively reduce on-chip storage consumption through MBF (Mapping Bloom Filter), while the latter improves forwarding efficiency by establishing different PITs for different interfaces. However, although the algorithm involved in MaPIT has improved the processing speed of PIT, it also needs more additional storage space, and the DiPIT architecture has to find every Bloom Filter in information retrieval to obtain the interface, which adds extra retrieval time, and this method increases the inevitable retrieval error due to the Bloom Filter's flaws.

Paper [14, 15] links PIT with hash retrieval to improve search speed and reduce storage consumption. Based on this method, the paper [16] proposes a modified MBF-based PIT storage architecture. The architecture uses a hash function to implement multiple hash mappings to improve the retrieval speed, and uses bitmap to realize the dynamic allocation of the address offset of the element memory unit, but as a static storage, hash table may cause a waste of storage space, and hash collisions can also affect the lookup performance.

To meet the requirements of processing of efficiency of PIT, this paper proposes an architecture of PIT based on hot table and implements further functional refinement of the PIT. The classification of interest is realized in the table and the matching rate of the PIT is satisfied. Thus efficient and accurate data query and forwarding can be realized in NDN.

The innovations of this paper are as follows:

- (1) A new type of PIT architecture based on hot table is designed. Interest packets that are about to enter the PIT are recorded and filtered, and a separate “priority channel” is established for the information that often passes the PIT.
- (2) The concept of “prefix value” is proposed in the PIT, and the information is graded according to the value of the prefix. Thus, it is ensured that the interest packet containing the higher value prefix can be queried and processed preferentially.
- (3) The self-maintenance strategy of the hot table is proposed, which realizes the dynamic adjustment of the data inflow direction. This strategy can not only ensure that the “priority channel” deals with the interest of higher value in the recent period, but also effectively prevents the overflow of the PIT.

2 Improvement of the Pending Interest Table and Related Design

2.1 Improvement of the Pending Interest Table

In the NDN architecture, the PIT provides two main functions, namely aggregation of interest packets and forwarding of data packets [12]. The same interest packets from different interfaces will be merged in the PIT, only the first arriving interest packet is routed to the potential responder; when the data packet arrives at the PIT, the incoming interface will be obtained from the PIT and passed out from the interface [6].

The PIT is usually very large, and for large NDN networks, it may reach the scale of several million to tens of millions [15]. In the 10G bandwidth environment, we assume a bad case that the average size of each interest packet and data packet is 64bytes. If the average round trip time per packet was 80 ms, PIT would need to contain about 800,000 entries. But in the case of bad network condition, the average round trip time of each packet will be greatly increased. Assuming that the average round trip time is 1 s, the PIT will contain about 10 million entries, PIT processing time for each packet need to reach 100 ns, this is just one link, on multiple link requesting conditions at the same time, the PIT will need a larger capacity and higher rate of access. Although memorizer like SRAM can guarantee the forwarding performance of PIT on a low bandwidth environment, but with the link rate increases, the memorizer capacity required by PIT becomes larger and larger and cannot meet the demand. Therefore, it is very important to improve the forwarding efficiency of PIT. In response to solve this problem, this paper proposes a PIT based on the hot table, as shown in Fig. 1.

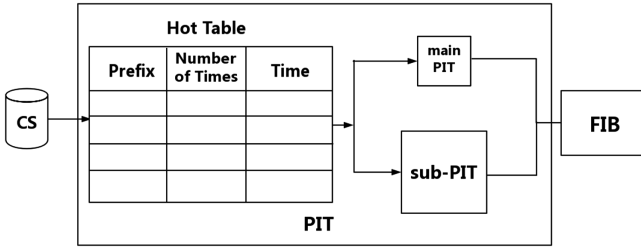


Fig. 1. The architecture of PIT based on hot table

In Fig. 1, the new PIT is composed of a hot table, a main PIT, and a sub-PIT. The hot table is located at the forefront of the entire PIT as an “indicator” for data shunting. The hot table is completely transparent to the data stream. It does not make any modification to the interest packet or data packet and only records the data, besides, it only records the prefix and quantity of the incoming information, it does not record the specific information, and has minimal demand for the storage space. By concentrating the high-value prefixes into the hot table for quick matching (the value determination method is described in detail in the second part), the pressure of all the information in the original architecture stored in one table for matching can be effectively alleviated. Also, by properly allocating data traffic, load balancing can be effectively achieved.

The function of the hot table is to store the higher-valued prefixes that are frequently requested within a certain period, and then perform centralized and fast processing on the information including these prefixes. The increase in the hot table does increase some of the cost compared to the original PIT architecture. However, since the hot table is small, typically store only a few dozen pieces of information, and only the content name prefix is stored in the table instead of the content. In PIT, since every forwarded content is recorded, PIT usually stores hundreds of thousands or even millions of messages. So the search cost brought by the hot table itself is negligible for the whole PIT.

The original PIT architecture is divided into two parts, namely the main PIT and the sub-PIT. The main PIT is responsible for storing the interest in the hot table and with high request frequency. The sub-PIT is responsible for storing the prefix that cannot match any of the hot tables and interest that the request frequency is low. In addition to the storage space, the internal architecture of the main PIT and the sub-PIT are consistent with the original PIT architecture.

In the entire network, because the popularity of the data is consistent with the Zipf distribution, that is, 20% of the content satisfies 80% of the request volume of the entire network, and the remaining 80% of the content satisfies the user’s 20% request volume [17]. Therefore, in the capacity allocation of the main PIT and the sub-PIT, we set the main PIT capacity to 20% of the maximum number of interest packets that the node may pass, and the sub-PIT capacity is 80% of the maximum number of interest packets that the node may pass.

The hot table consists of three items:

- (1) Prefix, the prefix is responsible for recording the prefix name of the incoming interest packet. We store the first-level prefix of the interest packet by default. When the proportion of entries in the main PIT is greater than a preset threshold, the second-level prefix is stored in the hot table. In this way, the total number of entries in the main PIT can be dynamically adjusted to prevent overflow of the main PIT and ensure efficient operation of the entire system.
- (2) The number of times, recording the total number of times the interest packet containing a prefix arrived.
- (3) Time, recording the last arrival time of interest packet containing a prefix.

2.2 Prefix Value Determination of Hot Table and Its Self-maintenance Strategy

Prefix Value Determination Method of Hot Table

Each prefix in the hot table has its value attribute. For a prefix, the value of each prefix can be given by combining the last access time of the prefix with the total number of times it is requested. The higher the value, the higher the popularity of the prefix, which is frequently accessed soon. The hot table calculates the value of each prefix in the table at regular intervals, and deletes the prefix that does not satisfy the preset threshold. Paper [18] uses the normalization method to obtain the popularity of data in the node storage space in the NDN, and then selectively replace the content. Here we learn from the method to get the value of each prefix.

T_{last} is the current time when a prefix was last accessed; $T_{current}$ is the current time; $T_{interval}$ is the time interval.

$$T_{interval} = T_{current} - T_{last} \quad (1)$$

R_{all} is the total number of times the prefix was requested. T_{first} is the first time the prefix is accessed, the average access interval ($T_{average}$) for the prefix is:

$$T_{average} = (T_{last} - T_{first})/R_{all} \quad (2)$$

The smaller the $T_{average}$, the more times the data is accessed in a short period.

T_{i_max} represents the maximum data in all $T_{interval}$; T_{i_min} represents the minimum in all $T_{interval}$, normalize $T_{interval}$:

$$T_{i_new} = \frac{T_{interval} - T_{i_min}}{T_{i_max} - T_{i_min}} \quad (3)$$

T_{a_max} represents the maximum of all $T_{average}$; T_{a_min} represents the minimum of all $T_{average}$, normalize $T_{average}$:

$$T_{a_new} = \frac{T_{average} - T_{a_min}}{T_{a_max} - T_{a_min}} \quad (4)$$

Thus each prefix value is:

$$P_{value} = \frac{1/(T_{a_new} + 1)}{T_{i_new}} \quad (5)$$

Self-maintenance Strategy of the Hot Table

To ensure the accuracy and timeliness of the data distribution of hot table, it is necessary to periodically calculate the value of each prefix for the hot table, and delete the prefix that does not meet the preset threshold. Here we remove the prefix with a value significantly smaller than the other prefixes according to the Pauta Criterion (PC). The Pauta Criterion is to assume that a set of test data only contains random error, calculate it to obtain the standard deviation, and determine an interval according to a certain probability, any data exceeding this standard deviation should be eliminated. Define each prefix value in the hot table as $x_1, x_2 \dots x_n$. Then, the average is:

$$\mu = (x_1 + x_2 + \dots + x_n)/n \quad (6)$$

The residual is:

$$g_i = x_i - \mu \quad (7)$$

According to the Bessel Formula, the standard deviation of these prefix values is:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n g_i^2} \quad (8)$$

The Pauta Criterion considers the probability of the value to be distributed in $(\mu - \sigma, \mu + \sigma)$ is 0.6827; the probability of the value to be distributed in $(\mu - 2\sigma, \mu + 2\sigma)$ is 0.9544 [19]. From this, it can be considered that the value of the prefix value is mostly concentrated in $(\mu - 2\sigma, \mu + 2\sigma)$. The probability of exceeding this range is only less than 4%. Since we only need to remove prefixes with a small value, we only need to remove prefixes with a value less than $\mu - 2\sigma$ for each self-maintenance.

Through the self-maintenance strategy of the hot table, the ‘‘cache pollution’’ effect of the historical data on the hot table can be prevented. Content prefixes that are frequently requested soon can be successfully retained in the hot table, and content prefixes with significantly reduced requests in the near future will be rejected, thus ensuring the freshness of the hot table and the timeliness of the entry in the hot table.

Routing and Forwarding Process

The NDN processing flow after improving the PIT architecture is shown in Fig. 2. Suppose a request for [youtube.com/music/jay/a.mp3](https://www.youtube.com/music/jay/a.mp3) arrives, first access the Content

Store (CS) for the newly arrived interest packet, and the CS is responsible for storing the data that has been requested by the router. If the data corresponding to the interest packet is queried in CS, the data is directly returned without subsequent operations; if there is no matching data, the interest packet is sent to the PIT for matching.

After entering the PIT, first check if there is a first or second level prefix (youtube.com or youtube.com/music) of this interest packet in the hot table. If it exists, the number of times of the prefix is incremented by one, and change the time corresponding to this prefix in the hot table to the current time. Then, the request is moved into the main PIT for matching; if no match exists, check to see if the hot table is full. If there is space in the hot table, put the prefix of the interest packet in the hot table, set the number of records to 1, set the time to the current time, and move the interest packet into the main PIT for matching. If the hot table is full, the table is full of prefixes that have been requested frequently and recently accessed, and the interest packet is moved to the sub-PIT.

When the data packet corresponding to the request returns, query it in the hot table first. If the prefix is in the hot table, the data packet enters the main PIT to find the corresponding information; otherwise, it goes into the sub-PIT for matching.

With this design, it is possible to ensure that a small amount of information with a higher value can enter the main PIT for fast matching; and for a large amount of information with a small value, it is moved into the sub-PIT for normal matching. Therefore, the priority query and processing of the high value interest can be effectively ensured, thereby improving the forwarding efficiency of the entire PIT.

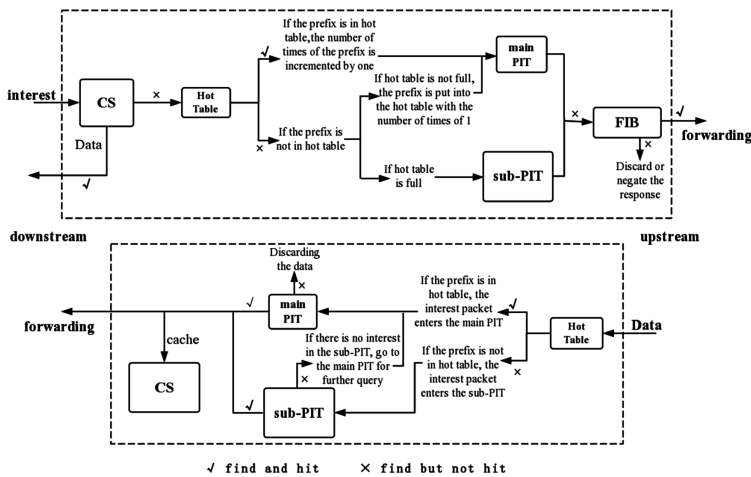


Fig. 2. NDN processing flow of PIT architecture with hot table

In response to a possible sudden outbreak of network traffic, we also added a flow control mechanism to the main PIT. When the number of entries in the main PIT reaches 80% of the capacity of the main PIT, the self-maintenance policy of the hot table is executed immediately, remove the prefixes that are less than $\mu - 2\sigma$ to make

room for the hot table. The subsequent interest packet stores only its second-level prefix in the hot table and replaces the original first-level prefix. The hot table follows the longest prefix matching principle, so the traffic under the original first-level prefix will partially flow into the sub-PIT, thus controlling the traffic entering the main PIT.

If the number of entries in the main PIT is more than 80% of the table capacity in at least two consecutive self-maintenance, the secondary hot table self-maintenance policy is executed. According to the Pauta Criterion, the probability that the prefix value in the hot table is distributed in $(\mu - \sigma, \mu + \sigma)$ is 0.6827. In this case, the prefix with a value less than $\mu - \sigma$ in the hot table is deleted, which provides more space for the hot table to store the newly arrived second-level prefix.

When the hot table self-maintenance time is reached, if the number of entries in the main PIT is less than 80%, the regular hot table self-maintenance policy is executed, and the second-level prefix is changed back to the storage first-level prefix in the hot table. In this way, we can prevent the occurrence that the entries in the hot table are filled with useless prefixes due to the accidental occurrence of sudden large flow. The flow control process of the PIT is shown in Figs. 3 and 4.

Before the hot table self-maintenance strategy is executed, the prefix youtube.com is recorded in the hot table, so all interests under the first-level prefix youtube.com are entered into the main PIT for matching.

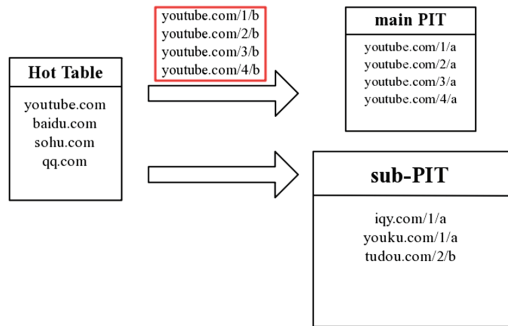


Fig. 3. Before the self-maintenance policy of the hot table is executed

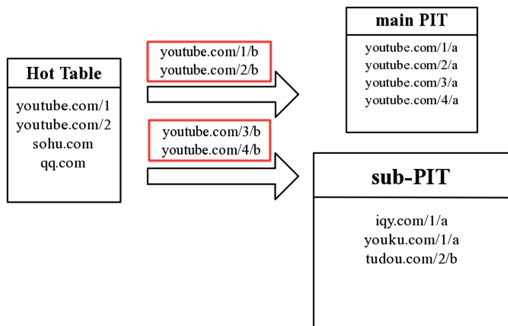


Fig. 4. After the self-maintenance policy of the hot table is executed

When the number of entries in the main PIT increases to the set threshold, the hot table removes some prefixes in the table through the self-maintenance policy and the hot table changes to only store the first-level prefix of the incoming interest, and then interest from later YouTube sites saved their second-level prefixes `youtube.com/1` and `youtube.com/2` in the hot table and removed the original first-level prefixes. When a new interest arrives in the hot table, the longest prefix matching rule is followed in the hot table, and all the interests that do not belong to `youtube.com/1` or `youtube.com/2` under `youtube.com` are diverted to the sub-PIT.

This method works better in the case of a large number of data requests (for example, video requests) in a short period. Because in this case the traffic is concentrated in several or dozens of prefixes, our hot table self-maintenance strategy is based on the last access time of prefix and the maximum of the prefix's average access time interval to do normalization processing. So in this case, the differentiation of different prefix values will be more obvious than random traffic, and more space will be spared to store the new second-level prefix, so it is more conducive to the flow control of the PIT.

In addition, we also consider that in the process of data forwarding, there may be a case that after a prefix in the hot table is deleted, the data packet will go to the sub-PIT for query because it cannot find the prefix in the hot table when it returns. In this case, we do the following: When the data packet returns, if the prefix exists in the hot table, it is directly queried to the main PIT. If there is no such prefix in the hot table, the sub-PIT is first queried, and then query it again in the main PIT if it does not exist in the sub-PIT. Since the main PIT is relatively small, it is possible to slightly increase the seek time while avoiding the problems that may exist in the foregoing, thereby improving the overall forwarding efficiency of the PIT.

3 Simulation Environment and Parameter Configuration

This chapter mainly conducts experimental analysis and performance evaluation. It is designed to analyze whether the improved PIT architecture can effectively improve the data forwarding efficiency in the data forwarding process of NDN.

3.1 Simulation Environment and Parameter Configuration

We use `ndnSIM` to implement the simulation of the PIT architecture. `ndnSIM` is an open source network simulation platform that can run on any available link layer protocol model. All NDN route forwarding experiments can be implemented on `ndnSIM` [20]. The experimental environment configuration table is shown in Table 1. Among them, the CPU uses Intel's 4-core processor `i5-4590` with 8 GB of memory, the system uses 64-bit `ubuntu12.04LTS`. The experiment uses four nodes as request nodes. The request rate ranges from 26000 packets per second to 66000 packets per second. The lifetime of the interest packet in the PIT is 1 s, and the size of the returned data packet is 1024 bytes.

Table 1. Experimental environment configuration table

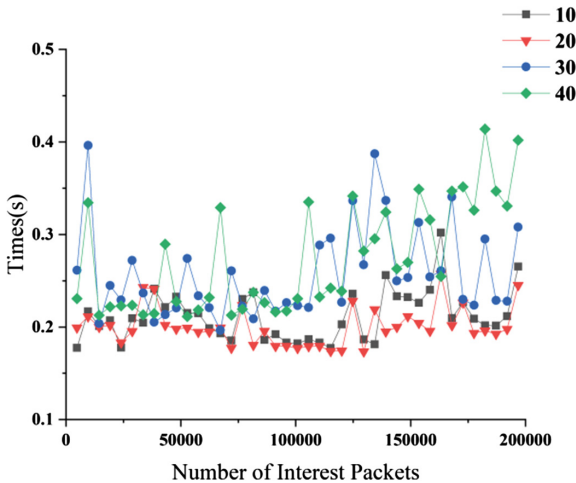
Main module	Specific configuration
CPU	Intel(R) Core(TM) i5-4590 (4 cores, Clock Speed: 3.30 GHZ)
RAM	8 GB
System	Ubuntu12.04
System digits	64-bit
ndnSIM version	2.3

3.2 Experimental Results and Analysis

Determine the Hot Table Capacity and Comparison of Round-Trip Time

Determine the Hot Table Capacity

The capacity of the hot table directly determines the forwarding performance of the whole PIT. Excessive hot table capacity not only increases the search burden, but also reduces the shunt effect of the interest packet; too small hot table capacity cannot maximize the performance of the PIT. To determine the most suitable hot table capacity, we made four nodes send a total of 200,000 interest packets as input data; then, every 400 packets were recorded from the time of sending interest packets to the time of receiving the corresponding data packets to test PIT performance under different hot table capacity.

**Fig. 5.** Comparison of round-trip time for different hot table capacities

In the experiment, by changing the capacity of the hot table, the round-trip time of each packet under different hot table capacities is obtained as shown in Fig. 5. The experimental results show that the round-trip time at the capacity of 30 and 40 is significantly greater than that at the capacity of 10 or 20, and the delay is the highest

and the image is the most dispersed when the capacity is 40. When the hot table capacity is 20, that is, when the hot table can store up to 20 entries, the overall round-trip time is significantly smaller and the image is the most stable. Therefore, when the number of user nodes is 4, the capacity of the hot table is set to 20, and the time required for the user to send the interest packet to obtain the data packet is the least, and the PIT can present its optimal performance.

Comparison of Round-Trip Time

The time elapsed by the requester from sending the interest packet to receiving its corresponding data packet is called the round-trip time. The round-trip time is one of the most important indicators in NDN performance measurement. In the experiment, we ensured that the content requested by the requesting node follows the Zipf distribution. That is, 20% of the content satisfies 80% of the request volume of the entire network, and the remaining 80% of the content satisfies the user's request amount of 20%, and the total number of interest packets is 200,000. The experimental results are shown in Fig. 6.

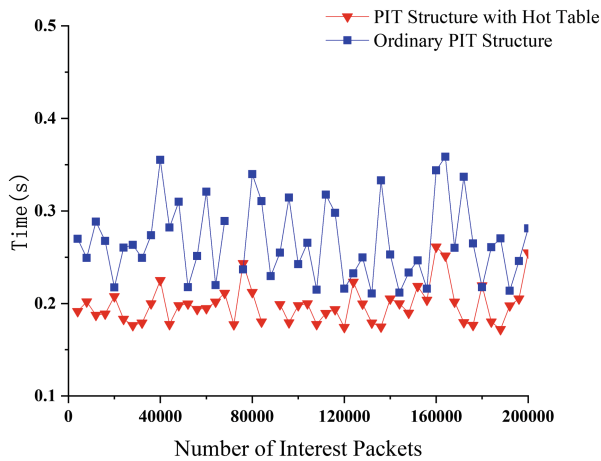


Fig. 6. Comparison of round-trip time

Figure 6 shows that the overall round-trip time of PIT architecture with a hot table is significantly less than that of conventional PIT architecture with the increase of the number of packets. During the same period, the round-trip time of the PIT architecture with the hot table is approximately 60%–80% of the original architecture, and at some point, it can reach less than 50% of the original architecture delay. The shunt of the hot table makes the matching of interest packets and data packets in the PIT more efficient, thus speeding up the round-trip time of each packet.

Comparison of Forwarding Rate

By measuring the forwarding rate under different PIT architectures are the most direct way to determine the performance of the PIT. To this end, we have measured the

forwarding rate of PIT architecture and conventional PIT architecture designed in this paper to compare their advantages and disadvantages. The experimental results are shown in Fig. 7.

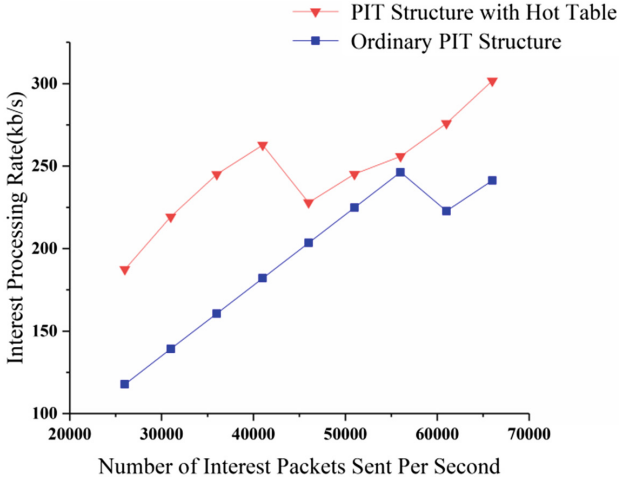


Fig. 7. Comparison of forwarding rate

In Fig. 7, the horizontal axis represents the number of interest packets sent per second, with the minimum of 26,000 interest packets sent per second and the maximum of 66,000 interest packets sent per second; the vertical axis represents the processing rate of the interest packets.

In the case of sufficient bandwidth, the more packets of interest packets are sent per second, the more packets that the PIT needs to process at the same time. Therefore, the processing rate of both architectures increases with the increase of the packet rate per second. According to Fig. 7, the PIT architecture with the hot table has a significantly higher forwarding rate than the original PIT architecture in the case of the same interest packet transmission rate.

Comparison of Packet Loss Rate

Here, the packet loss rate refers to the percentage of all interest packets in the PIT that are replaced by the replacement policy or the interest packets that have been deleted in the timeout period. Since the hot table is first queried during the return of the data packet, there may be a case that when a prefix in the hot table is deleted, the packet will go to the sub-PIT to query because it cannot be found in the hot table. Therefore, an experiment is needed to verify whether this situation will affect the forwarding performance of the PIT. Figure 8 shows the change in packet loss rate for different PIT architectures with increasing packet rate.

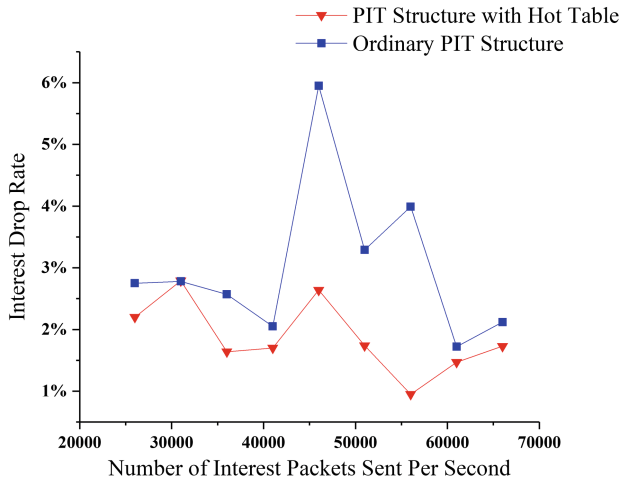


Fig. 8. Comparison of packet loss rate

The experimental results show that the PIT architecture with the hot table can maintain a lower packet loss rate than the conventional PIT architecture. The main reason is that the hot table based PIT architecture enables the interest packet or data packet to query the corresponding entry more quickly when it arrives so that it can be processed in time before the entry in the table times out.

4 Conclusion

Although the NDN has many advantages, there are some problems in the design of some details, such as the interest processing problem of the PIT. In this paper, the design requirements of fast processing and efficient forwarding of the PIT are realized. By improving the architecture of the PIT, we proposed the concept of “prefix value”, which records the high-value prefixes in PIT by using hot table, and makes the information containing these prefixes be queried and processed first when the interest packet enters and the data packet returns, thereby saving the overall processing time and ensuring the efficient operation of the NDN router. The experimental results show that the proposed scheme can effectively improve the PIT’s processing performance. However, the sample size of data collected in this paper is not very large. Also, in addition to the optimization design of the PIT, there is still room for further research and improvement in the collaborative cache of the PIT and the CS. We will further improve the performance and processing efficiency of NDN on the basis of this paper.

References

1. Zhang, L., et al.: Named data networking (NDN) project. In: Relatório Técnico NDN-0001, pp. 157–158. Xerox Palo Alto Research Center-PARC, Los Angeles (2010)
2. Lei, K.: Information Centric Networking and Named Data Networking. Peking University Press, Beijing (2015). (in Chinese)
3. Zhang, L., et al.: Named data networking. In: Proceedings of ACM SIGCOMM Computer Communication Review, vol. 44, pp. 66–73. ACM, New York (2014)
4. Perino, D., Varvello, M.: A reality check for content centric networking. In: Proceedings of ACM SIGCOMM, pp. 44–49. ACM, New York (2011)
5. Wang, Y., et al.: Scalable name lookup in NDN using effective name component encoding. In: Proceedings of 2012 IEEE 32nd International Conference on Distributed Computing Systems, pp. 688–697. IEEE, Macau (2012)
6. Dai, H., et al.: On pending interest table in named data networking. In: Proceedings of the Eighth ACM/IEEE Symposium on Architectures for Networking and Communications Systems, pp. 211–222. ACM, New York (2012)
7. Huang, H., Wang, Y., Lan, J.: An instant-triggered PIT item aging method. *J. Inf. Eng. Univ.* **16**(02), 152–156 (2015). (in Chinese)
8. Liu, D., Hu, Y., Zhuang, L.: Research on fast forwarding response mechanism in NDN. *J. Zhengzhou Univ.* **12**(2), 68–72 (2015). (in Chinese)
9. Alubady, R., Hassan, S., Habbal, A.: Adaptive interest lifetime in named data networking to support disaster area. *J. Telecommun. Electron. Comput. Eng. (JTEC)* **10**(2–4), 29–34 (2018)
10. Alubady, R., Hassan, S., Habbal, A.: HLLR: highest lifetime least request policy for high performance pending interest table. In: Proceedings of 2016 IEEE Conference on Open Systems (ICOS), pp. 42–47. IEEE, Langkawi (2016)
11. Li, Z., Bi, J., Wang, S., Jiang, X.: Compression of pending interest table with bloom filter in content centric network. In: Proceedings of 2012 ACM 7th International Conference on Future Internet Technologies, pp. 46. Springer, Heidelberg (2012)
12. Li, Z., Liu, K., Zhao, Y., Ma, Y.: MaPIT: an enhanced pending interest table for NDN with mapping bloom filter. *IEEE Commun. Lett.* **18**(11), 1915–1918 (2014)
13. You, W., et al.: DiPIT: a distributed bloom-filter based pit table for CCN nodes. In: Proceedings of 2012 21st International Conference on Computer Communications and Networks (ICCCN), pp. 1–7. IEEE, Munich (2012)
14. Yuan, H., Crowley, P.: Scalable pending interest table design: from principles to practice. In: Proceedings of 2014 IEEE INFOCOM Conference on Computer Communications, pp. 2049–2051. IEEE, Toronto (2014)
15. Varvello, M., Perino, D., Linguaglossa, L.: On the design and implementation of a wire-speed pending interest table. In: Proceedings of 2013 IEEE Conference on Computer Communications Workshops, pp. 369–374. IEEE, Turin (2013)
16. Xu, Y., et al.: Research on PIT storage structure of named data networking based on improved MBF. *J. Chongqing Univ. Posts Telecommun.* **30**(1), 61–67 (2018). (in Chinese)
17. Li, G.: Service popularity-based smart resources partitioning for fog computing-enabled industrial Internet of Things. *IEEE Trans. Ind. Inform.* **14**(10), 4702–4711 (2018)
18. Liu, Q., et al.: Popularity-based in-network cache replacement strategy in named data networking. *Comput. Eng. Appl.* **906**(11), 81–85 (2018). (in Chinese)
19. Zhang, X., Cheng, M., Xiao, F.: Using origin to compare data outliers. *Exp. Sci. Technol.* **10**(1), 74–76 (2012). (in Chinese)
20. Afanasyev, A., Moiseenko, I., Zhang, L.: ndnSIM: NDN simulator for NS-3. University of California, Los Angeles (2012)

Applications on Internet of Things



Real-Time Bridge Structural Condition Evaluation Based on Data Compression

Jingpei Dan^{1(✉)}, Ling Liu², Yuming Wang³, Junji Chen¹, and Xia Huang¹

¹ Chongqing University, Chongqing 400044, China
jingpeidan@cqu.edu.cn

² Georgia Institute of Technology, Atlanta 30332, USA

³ Huazhong University of Science and Technology, Wuhan 430074, China

Abstract. Detecting structural damage in real time is important and challenging for bridge structural health monitoring systems, especially when large amount of time series monitoring data are collected for continuous monitoring and evaluation of abnormal conditions. Conventional approaches fail to efficiently process such large-scale data in real time due to high storage and processing cost. In this paper, we present an efficient real-time bridge structural condition evaluation based on data compression. We introduce an efficient time series representation to compress sensor data into symbol streams by applying symbolic aggregate approximation (SAX), which transforms sensing data into symbolic representation to reduce dimension while preserving important features and guaranteeing low-bounding distance. Upon receiving sensing data in real time, we compress raw data into SAX representation before evaluation. Then, we evaluate bridge structural condition by performing classification based on compressed data efficiently. The proposed method is evaluated using a typical real bridge data set from SMC. Compared with the prediction results on original data using existing methods, our approach reduces the processing time from hours to several seconds with improved accuracy, showing that the proposed method is effective in improving both efficiency and accuracy of bridge structural condition evaluation in real time.

Keywords: Bridge structural condition evaluation · Symbolic aggregate approximation · Data compression · Time series

1 Introduction

Bridge structural health monitoring (BSHM) has been widely applied in real-time monitoring of large bridge with the advantages of uninterrupted traffic, real-

This research is supported by National Natural Science Foundation of China (No. 51608070), Chongqing general project of basic science and advanced technology (No. cstc2016jcyjA0022), USA NSF CISE SaTC grant (No. 1564097), an IBM faculty award, and Fundamental Research Funds for the Central Universities (No. 2019CDXYJSJ0021).

time, and full-time service. In conventional BSHM systems, there are hundreds of sensors of varying types installed on different parts of bridges to monitor structural response and environment changes, usually including vehicle load, girder strains, accelerations, wind, temperature, and humidity. Figure 1 shows that a typical BSHM system consists of a data acquisition subsystem that collects sensor data on the bridge and transmits data in batch to its bridge structural condition evaluation subsystem for anomaly detection [1]. Numerous types of sensor data are collected continuously when a BSHM system is in operation. Generally speaking, the size of collected data has been on the order of GBs per day and TBs per year and such rates are growing rapidly. As all measured data from the sensors are collected, stored and processed by BSHM on its remote servers. Supporting real-time damage identification has been a critical challenge for the next generation of the Bridge structural health monitoring systems and conventional schemes fail to support anomaly detection and identification in real time.

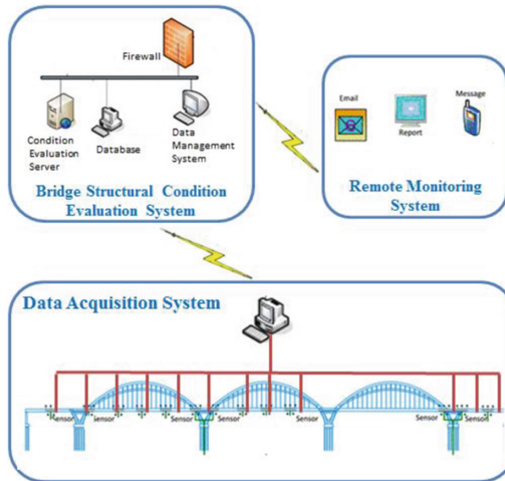


Fig. 1. Bridge structural monitoring system

Bridge structure deterioration is a gradually changing procedure of structural condition. When some slight damages are detected, it usually takes conservative maintenance, otherwise it will be too late for maintenance when it is evaluated as severe damaged. To avoid serious safety accident, it is important to detect the deteriorating change trend from slight damaged to severe damaged. That means compared to evaluating condition using data at a time point, it is more reasonable to use a time series data that reflects deteriorating change trend.

In recent years, in addition to the statistical model-based bridge structural damage detection methods [2,3], support vector machine [4,5], artificial neural networks [6,7], wavelet transform [8], convolutional neural network [9], and other

intelligent methods have been applied in BSHM [10]. However, it is difficult for the existing approaches to effectively process such kind of large and various data in real time and locally due to computational complexity.

To address above challenge, we develop an efficient real-time bridge structural condition evaluation method with data compression. Instead of conventional feature extraction, we introduce time series representation to compress real-valued sensor data into symbol streams by applying symbolic aggregate into approximation (SAX) [11]. SAX is a time series representation that performs well in various domains [12, 13]. It transforms real-valued data into symbolic representation to reduce dimension while remains important features and guarantee low-bounding distance as well. Taking these advantages, we preprocess and compress raw data into SAX representation ready for further evaluation. Then we evaluate bridge structural condition by classification based on the compressed data. We evaluate the effectiveness of our proposed approach on a real bridge data set, provided by the Structural Monitoring and Control Research Center (SMC) of Harbin Institute of Technology. Comparing the experimental results on original data and compressed data, the processing time is reduced from hours to several seconds and the accuracy is improved as well, demonstrating the effectiveness of the proposed method for improving efficiency and accuracy of bridge structural condition evaluation in real time.

The rest of the paper is organized as follows. The proposed bridge structural condition evaluating method is briefly introduced in Sect. 2; The detail of data compression is described in Sect. 3. To validate the proposal, the experiments and results are discussed in Sect. 4. The conclusion and future works are given in Sect. 5.

2 Bridge Structural Condition Evaluation Based on Data Compression

Our bridge structural condition evaluating method is composed of two major parts: data compression and data analysis. Data analysis is based on data compression, as explained in the following subsections.

2.1 Data Compression

Data compression is a common technique used for managing large amount of data to reduce storage cost and processing cost in many data-driven application services. When transforming a large amount of sensor data into a compressed representation, one important criterion is to maintain data utility during the compression process for data analysis. Thus, a main challenge of data compression is to compress data in a proper ratio to reduce dimension but still preserve the desired essential features.

To address this problem, we compress data adaptively in terms of its status. During the service period, a bridge is always in healthy state for a long period of time before damages occur. There is lots of redundancy in most of the monitoring

data that indicates healthy condition. We compress this class of sensor data using high compress ratio, while maintaining important features for damage detection. Meanwhile, the compressed data in healthy state are transmitted to the remote server and stored for future analysis. Since damage happens over time, some sensor data is gradually deviated from healthy state, and such data may contain more key features than baseline data. To keep these essential features for damage detection, we compress the data using low compress ratio and save the compressed data locally. This allows us to zoom into those sensor data for damage identification, which greatly reduces both storage and search space cost for sensor stream data processing.

A variety of data compression techniques have been proposed and applied in data mining algorithms, such as statistic models, principal component analysis, and wavelet transform are most commonly used methods. In order to predict the damage changing trend, it is more informative if data mining and analysis are performed based on time series data. This consequently requires us to consider using time series data representation to compress times series data. Among numerous time series data representations, we adopt SAX [11] as our data compression scheme in our bridge structural condition evaluation method (see Sect. 3 for detail).

2.2 Data Analysis

We combine data analysis with data compression such that the sensor data are transformed through compression with a given compression ratio (see more detail in Sect. 3).

Though many damage detection techniques have been proposed in recent years, they mainly concern on improving accuracy. To verify how efficiency our framework may improve over existing representative methods, it is fair to choose the commonly used baseline damage detection technique in this study. In addition, instead of data analysis on remote server as conventional approaches, we can perform this framework locally to realize nearly real-time detection.

K-NN is an example-based classification method which is one of the simplest machine learning algorithms and widely used in many areas. For fair comparison, we choose K-NN as evaluating algorithm. The basic idea of K-NN is to determine the class of the sample according to the category of K samples that are closest to it. It is important to apply an appropriate similarity measure to improve efficiency and accuracy of K-NN classification method. Euclidean Distance is one of the most widely used distance measurements suitable for real-valued time series due to its simplicity and low complexity. The Euclidean-distance between X and Y with both of length n is given in (1).

$$D(X, Y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

We employ the Euclidean Distance method to calculate distances among real-valued time series, and adopt the MINDIST function in (3) in SAX repre-

sentation to calculate the similarity measures among symbolic representations. Our design is extensible with respect to both classification algorithms for supervised learning and distance functions. We use the K-NN classifier as an example to illustrate the development of our approach.

3 Symbolic Based Data Compression

There are numerous time series representations for time series data compression. Commonly used methods include the Discrete Fourier Transform (DFT), the Discrete Wavelet Transform (DWT), the Piecewise Aggregate Approximation (PAA) and etc. The suitable choice of representation greatly affects the ease and efficiency of time series based data analysis. For example, wavelets have the useful multiresolution property, but are only defined for time series that are an integer power of two in length [11]. A common feature important to all above representations is real-valued data representation, which limits the algorithms, data structures and definitions available for them. This limitation inspired some developments of symbolic representations. Among them, Symbolic Aggregate Approximation (SAX) is one of the most popular representations with good performance in various applications. SAX not only reduces dimensions adaptively but also allows low-bounded distance comparison between data. In our framework, SAX is chosen as the compression method for compressing BSHM data prior to sending them for time series data analysis.

SAX is a piecewise-based symbolic presentation of time series, which allows a time series of length n to be expressed by a string of length w . Since it is meaningless to compare time series with different offsets and amplitudes, time series data is normalized with mean of zero and standard deviation of one before transforming it into the PAA representation [14].

PAA is an efficient time series dimensionality reduction method based on discretization. A time series of length n , denoted by $C = \{c_1, c_2, \dots, c_n\}$, can be represented by a w -dimensional vector, denoted as $\bar{C} = \{\bar{c}_1, \bar{c}_2, \dots, \bar{c}_w\}$. The i th element of \bar{C} is calculated as follows:

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}*(i-1)+1}^{\frac{n}{w}*i} c_j \quad (2)$$

PAA approach to represent time series data is simple and intuitive, and can significantly reduce the dimension of time series data, and yet has been shown to be competitive with those more sophisticated dimensionality reduction techniques, like Fourier transforms and wavelets.

Since the normalized time series have a highly normal distribution, it is easy to set breakpoints that will produce an equal-sized area under Gaussian curve, and use same symbol to represent PAA coefficients that are distributed in the same range. Then the PAA representation of time series is transformed into a string, which called the SAX representation and \hat{c} is used to indicate the SAX. Table 1 gives the breakpoints β_i when a is in the range of 3 to 10.

Table 1. Breakpoints based on different values of α .

β_i	3	4	5	6	7	8	9	10
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15	-1.22	-1.28
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67	-0.76	-0.84
β_3		0.67	0.25	0	-0.18	-0.32	-0.43	-0.52
β_4			0.84	0.43	0.18	0	-0.14	-0.25
β_5				0.97	0.57	0.32	0.14	0
β_6					1.07	0.67	0.43	0.25
β_7						1.15	0.76	0.52
β_8							1.22	0.84
β_9								1.28

Figure 2 shows the process that transforms a time series to a SAX representation. A time series of length 128 is firstly discretized to obtain a PAA approximation of length 8, then set $\alpha = 3$, map the PAA into SAX symbols via predetermined breakpoints shown in Table 1. Thus, the time series is mapped to the string $\hat{c} = baabccbc$.

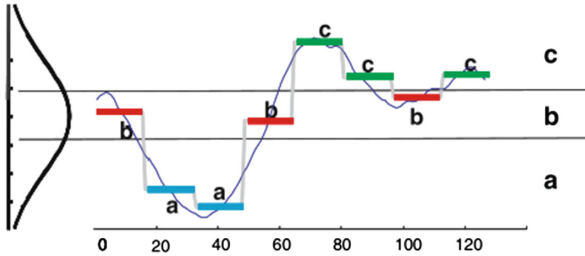


Fig. 2. A brief illustration of the process of symbolizing a time series to a character string

The MINDIST function defined in (3) is used to measure the distance between two SAX strings. The sub function $dist()$ is formed by Table 1. Table 2 shows the distance between two symbols when $\alpha = 4$, and the value in the cell(r, c) can be calculated by (4).

$$MINDIST(\hat{Q}, \hat{C}) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (dist(\hat{q}_i - \hat{c}_i))^2} \tag{3}$$

$$cell(r, c) = \begin{cases} 0 & if |r - c| \leq l \\ \beta_{max(r,c)-1} - \beta_{min(r,c)} & otherwise \end{cases} \tag{4}$$

Table 2. The distance between two symbols when $\alpha = 4$.

	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>
<i>a</i>	0	0	0.67	1.34
<i>b</i>	0	0	0	0.67
<i>c</i>	0.67	0	0	0
<i>d</i>	1.34	0.67	0	0

The distance measure method satisfies the lower bounding. The lower bounding feature is particularly exciting because it allows one to run certain data mining algorithms on the efficiently manipulated symbolic representation, while producing identical results to the algorithms that operate on the original data.

In our framework, the original sensor data are compressed by applying SAX representation. Compression ratio can be adaptively set in terms of initial evaluating result. The damage degree evaluation results are sensitive to the settings of data compression ratio. In this study, we use experimentation to set the compression ratios.

4 Experimental Evaluation on SMC Real Bridge Dataset

In order to validate the effectiveness of our bridge structural condition evaluation method, we conduct a set of experiments on a dataset: SMC real bridge dataset. All experiments are conducted on a computer server with 3.30 GHz CPU, 16 GB main memory and running on a Windows operating system. The experimental results show that our proposed method performs well on SMC real bridge dataset with respect to runtime performance and accuracy. The experimental setup and result analysis are elaborated as follows.

4.1 SMC Real Bridge Dataset

In order to verify the effectiveness of the proposal in real application, an experiment has been carried out on the BSHM dataset provided by the Structural Monitoring and Control Research Center (SMC) of Harbin Institute of Technology [15].

The bridge is precast concrete double tower cable-stayed bridge with the main span of 260 m. The bridge is 510 m long and 11 m wide. After a period of working years, cracks are observed at the bottom of the mid-span girder. A SHM system has been implemented after the bridge is repaired and reopened for traffic to monitor and acquire time series data. Many BSHM methods based on Generalized Regression Neural Network, PCP, and so on all have good performances on this dataset [16–18].

As part of the SHM system, fourteen uniaxial accelerometers are installed on the deck, and one biaxial accelerometer is attached to the top of the south tower. The layout of acceleration sensors is shown in Fig. 3.

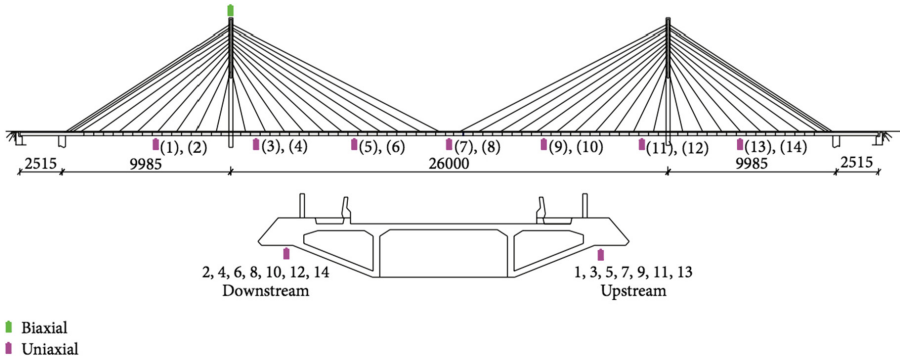


Fig. 3. Acceleration sensors settings.

The acceleration data are acquired by the SHM system from January 1 to July 31, 2008. The sampling frequency of the acceleration is 100 Hz. Data during twelve days are selected to represent the benchmark time history of the bridge, from healthy status to damaged status. Additional accelerometers are implemented in August 7 to August 10, 2008 to collect vibration data two hours a day. These data are used to represent the final damage status. After normalized the experimental data, sliding window is applied as well. The window size is set as 3300 and run-length is 1320. Figure 4 shows the graphic of the bridge structure acceleration time series data.

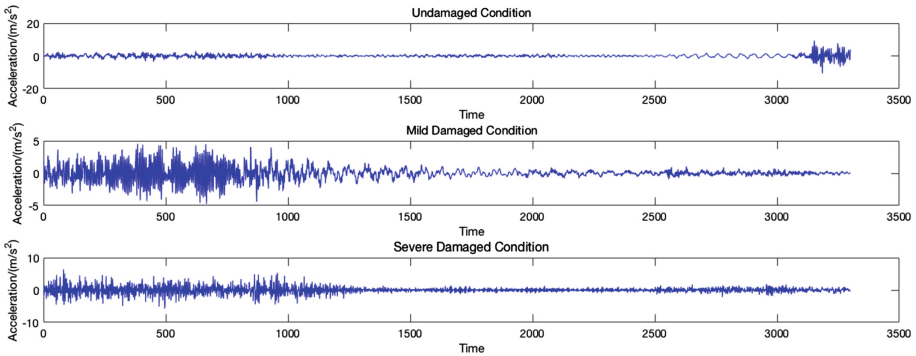


Fig. 4. Healthy, mild, and severe damaged time series data.

4.2 Experimental Results and Analysis

We conduct experiments on SMC bridge dataset to verify the effectiveness of our proposal in terms of evaluation accuracy and time cost. Since K-NN is the

most common used classifying algorithm, it is fair to compare the performances of different situation of data, i.e., uncompressed and different compressing algorithms, by adopt K-NN to classify structural conditions. In addition, SVM is also chosen to perform classification as an intelligent method. The experimental results are analyzed respectively as follows. Firstly, we apply K-NN as classifying algorithm, the length of the original acceleration time series of the bridge is 3300, and the compression ratio is set as 20, the data length of PAA representation and SAX representation is 165. Figure 5 shows the process of the time series with the length of 200 transformed into SAX representation. The average accuracy and efficiency are shown in Table 3. The consumption of time is reduced from about 2 h to 14 s, at the same time, the average accuracy is increased from 77% to 93%. In addition, SAX performs better than PAA. It is indicated that the structural condition assessment method based on SAX significantly better than the traditional method in terms of time cost and evaluation efficiency.

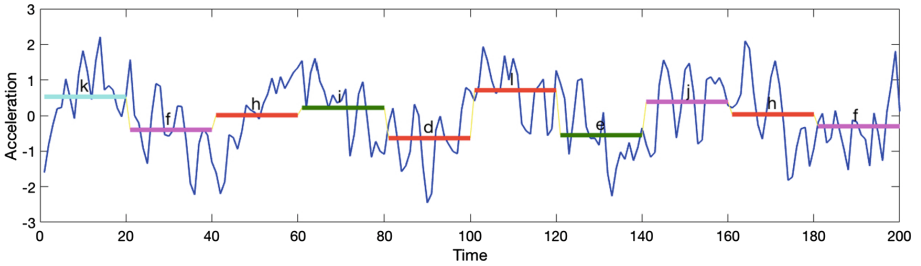


Fig. 5. SAX representation of time series data.

Table 3. Comparative experimental results by applying K-NN

	Average accuracy (%)	Average time cost (second)
Raw Data	51.68	1517
PAA Data	75.13	5.06
SAX Data	81.00	4.79

Then, we adopt SVM to classify structural conditions. Radial basis function is set as the kernel function of SVM. As shown in Table 4, the consumption of time is reduced from about 3.8 h to 17 s, while the average accuracy is increased from 79.12% to 91.2%. In this case, the proposed method based on SAX is also validated to have the best performances both in time cost and average accuracy.

Comparing above two experiments, classifying algorithms have little influences on the evaluation accuracy, but the time cost is increased by using more

Table 4. Comparative experimental results by applying SVM

	Average accuracy (%)	Average time cost (minute)
Raw data	79.12	227
PAA data	89.28	0.36
SAX data	91.2	0.29

complicated classifying algorithms. In addition, no matter which classifying algorithm is applied, compressing raw data by using SAX can improve efficiency and greatly reduce time cost. The experimental results indicate that our bridge structural condition evaluation based on SAX has good performance in real application.

5 Conclusion

Real time bridge structural condition evaluation represents an important class of analysis as a service. In this paper, we have presented a real-time bridge structural condition evaluation method based on symbolic data compression. We introduce time series representation to compress real-valued sensor data into symbol streams by applying SAX. SAX transforms real-valued data into symbolic representation to reduce dimension, meanwhile essential features can be remained and low-bounding distance can be guaranteed as well. Sensing data in real-values are preprocessed and compressed into SAX representation for subsequent structural condition evaluation. Then, we evaluate bridge structural condition by classifying compressed data. We have validated the effectiveness of the proposed approach with experiments performed on SMC real bridge dataset. We have compared the experimental results performed on original data, PAA, and SAX. The experimental results show that the processing time is reduced from hours to several seconds while improving the accuracy by our method. We also compare the performances of different classifying algorithms. The results show that the accuracy is not much affected by classifying algorithm. It is indicated that our proposal is effective to evaluate bridge structural condition with high efficiency and accuracy in real time.

References

1. Farrar, C.R., Worden, K.: An introduction to structural health monitoring. *Philos. Trans. R. Soc. Lond.* **A365**(1851), 303–315 (2007)
2. Gul, M., Catbas, F.N.: Statistical pattern recognition for Structural Health Monitoring using time series modeling: theory and experimental verifications. *Mech. Syst. Sig. Process.* **23**(7), 2192–2204 (2009)
3. Yao, R., Pakzad, S.N.: Autoregressive statistical pattern recognition algorithms for damage detection in civil structures. *Mech. Syst. Sig. Process.* **31**, 355–368 (2012)

4. Hasni, H., Alavi, A.H., Jiao, P., Lajnef, N.: Detection of fatigue cracking in steel bridge girders: a support vector machine approach. *Arch. Civ. Mech. Eng.* **17**(3), 609–622 (2017)
5. Huynh, C.P., Mustapha, S., Runcie, P., Porikli, F.: Multi-class support vector machines for paint condition assessment on the Sydney Harbour Bridge using hyperspectral imaging. *Struct. Monit. Maint.* **2**(3), 181–197 (2015)
6. Mehrjoo, M., Khaji, N., Moharrami, H., Bahreininejad, A.: Damage detection of truss bridge joints using Artificial Neural Networks. *Expert Syst. Appl.* **35**(3), 1122–1131 (2008)
7. Arangio, S., Bontempi, F.: Structural health monitoring of a cable-stayed bridge with Bayesian neural networks. *Struct. Infrastruct. Eng.* **11**(4), 575–587 (2015)
8. Xia, Y.-X., Ni, Y.-Q.: A wavelet-based despiking algorithm for large data of structural health monitoring. *Int. J. Distrib. Sens. Netw.* **14**(12), 1550147718819095 (2018)
9. Tang, Z., Chen, Z., Bao, Y., et al.: Convolutional neural network-based data anomaly detection method using multiple information for structural health monitoring. *Struct. Control Health Monit.* **26**, e2296 (2018)
10. Worden, K., Manson, G.: The application of machine learning to structural health monitoring. *Phil. Trans. R. Soc. Lond.* **A365**(1851), 515–537 (2007)
11. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: *Proceedings of Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2001)*. ACM SIGMOD Press, pp. 2–11 (2003)
12. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. *Data Min. Knowl. Disc.* **15**(2), 107–144 (2007)
13. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: a survey. *IEEE Trans. Knowl. Data Eng.* **26**(9), 2250–2267 (2014)
14. Keogh, E., Lonardi, S., Chiu, B.Y.C.: Finding surprising patterns in a time series database in linear time and space. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, pp. 550–556. ACM SIGKDD Press (2002). <https://doi.org/10.1145/775047.775128>
15. Li, S., Li, H., Liu, Y., Lan, C., Zhou, W., Ou, J.: SMC structural health monitoring benchmark problem using monitored data from an actual cable-stayed bridge. *Struct. Control Health Monit.* **21**(2), 156–172 (2014)
16. Yang, Y., Nagarajaiah, S.: Blind denoising of structural vibration responses with outliers via principal component pursuit. *Struct. Control Health Monit.* **21**(6), 962–978 (2014)
17. Gharibnezhad, F., Mujica, L.E., Rodellar, J.: Applying robust variant of Principal Component Analysis as a damage detector in the presence of outliers. *Mech. Syst. Sig. Process.* **50–51**, 467–479 (2015)
18. Huang, Y., Beck, J.L., Wu, S., Li, H.: Bayesian compressive sensing for approximately sparse signals and application to structural health monitoring signals for data loss recovery. *Probabilist. Eng. Mech.* **46**, 62–79 (2016)



Task Assignment Algorithm Based on Social Influence in Mobile Crowd Sensing System

Anqi Lu and Jinghua Zhu(✉)

Heilongjiang University of Science and Technology, Harbin, China
luanqi960613@163.com, zhujinghua@hlju.edu.cn

Abstract. In the mobile crowd sensing (MCS) system, task assignment is a core and common research issue. Based on the traditional MCS platform, there is a cold start problem. This paper introduces social networks and communication networks to solve the cold start problem. Therefore, this paper draws on social influence to propose a greedy task assignment algorithm H-GTA. The core idea of the algorithm is to first use the communication network to select seed workers in a heuristic manner according to the recruitment probability and then the seed workers spread the task on social networks and communication networks simultaneously. The publisher selects workers to assign the task in a greedy way to maximize the task's spatial coverage. When calculating the probability of recruitment, this paper considers various factors such as worker's ability, stay time and worker's movement to improve the accuracy of recruitment probability. Considering the influence of worker's movement on recruitment probability, a worker's movement prediction algorithm based on meta-path is proposed to analyze worker's movement. The experimental results show that compared with the existing algorithms, the algorithm in this paper can guarantee the time constraint of the task, and have better performance in terms of spatial coverage and running time.

Keywords: Mobile crowd sensing · Social influence · Task assignment · Time constraint · Spatial coverage · Meta-path

1 Introduction

With the development of the Internet of Things, mobile Internet and cloud computing, there are now various ways to collect city information. Among them, the popularity of mobile devices, and the growing demand for smart sensing in the city provide a method for urban sensing, called Mobile Crowd Sensing (MCS) [1]. MCS has become a kind of new way to collect real-time environmental information by using mobile workers' smart devices. This paper uses spatial coverage as a measure of sensing quality for environmental monitoring tasks.

Most of the existing methods for task assignment are based on a specialized MCS platform. Assumed that there is a large pool of workers, a subset

of workers is selected based on optimization objectives and constraints. However, in some cases, this assumption is not true, as the platform has just been established, so-called cold-start problem. This paper proposes a task assignment algorithm for mixing social networks and communication networks. Through the communication network, the task can be pushed to the workers near the task release location, and then the task is propagated through the social network and communication network to solve the cold start problem.

The task publisher issues a sensing task at a certain location on the communication network. First, during the initial propagation period of the task, the communication network is used to push the task to the worker near the task issued location, and a part of the workers is selected as the seed set, and then during the additional propagation period of the task, the workers in the seed set use social networks and communication networks to push tasks to their friends or neighbors, and a part of the workers is selected and the workers in the seed set are grouped to form the final recruited workers set. The goal of the algorithm proposed in this paper is to select workers to assign the task according to the recruitment probability to maximize the spatial coverage of the MCS task under the constraints of the number of seeds and the number of final recruited workers.

When considering the probability of recruitment, this paper considers various factors such as worker's ability, stay time and worker's movement from the perspective of the task publisher to improve the accuracy of calculating the recruitment probability. When considering the impact of worker's movement on the calculation of recruitment probability, an offline worker's movement prediction method based on meta-path is proposed in this paper.

The main contributions of this paper are as follows:

- We propose a time-constrained task assignment algorithm for mixing social networks and communication networks to solve the cold start problem of traditional MCS platforms.
- We propose a method of calculating the recruitment probability, considering worker's ability, incentive mechanism, stay time, the number of influenced times, and worker's movement to improve the accuracy of calculating the probability.
- We propose a worker's movement prediction method based on the meta-path to calculate the worker density in the sensing area.
- We perform the experiments to verify the time constraint and the performance of the proposed method in terms of spatial coverage and running time.

2 Related Work

2.1 MCS Task Assignment

The MCS task assignment algorithms can be divided into two categories, one is to maximize the sensing quality with certain constraints. Wang et al. [2] propose to solve the MCS worker recruitment assisted by the influence propagation on the social network, and its goal is to maximize the temporal-spatial coverage

without exceeding the incentive budget. Zhang et al. [3] study the selection of a subset of mobile workers to maximize coverage quality in the case of budget constraints. The other is to minimize the cost with sensing quality constraints. Karaliopoulos et al. [4] solve the worker recruitment problem under opportunistic networks in the context of mobile crowdsourcing to collect location-based data and minimize total cost while ensuring total coverage of PoI. Xiong et al. [5] propose a new MCS framework that considers worker’s privacy under the constraints of minimizing the number of task assignment requirements and sensing area coverage, to reduce worker’s energy consumption and data transmission caused by MCS task assignment and data collection.

2.2 Influence Maximization on Social Networks

The problem of influence maximization was first proposed by Domiggos and Richardson et al. [6] who modeled the problem as a Markov random field and solved the problem with a heuristic algorithm. Kempe et al. [7] design a greedy algorithm, they first introduce independent cascade (IC) model and linear threshold (LT) model into the problem, which is of great enlightenment to the later research. Li et al. [8] propose a new network model and influence propagation model, which considers the influence propagation of the online social network and the physical world.

The method proposed in this paper borrows the idea of influence maximization on social networks, but it differs a lot from the above research in the following aspects:

- (1) *Different influence propagation model.* This paper considers the influence propagation between friends and neighbors.
- (2) *Different optimization goal and constraint.* This paper limits the number of seeds and final recruited workers in order to maximize the task’s spatial coverage.
- (3) *Different seeds selection algorithm.* When selecting a seed node, this paper considers not only its influence but also the increase of the spatial coverage.

3 Network Model and Problem Definition

3.1 Network Model

Definition 1 (Social Network). $G_S = \langle W, E_f \rangle$, where W denotes the set of workers, E_f denotes the edge set of workers’ friend relationship.

Definition 2 (Communication Network). $G_C^t = \langle W, E_n^t \rangle$, where W denotes the set of workers, E_n^t denotes the edge set of workers’ neighbor relationship. $w_i.x_t$ and $w_i.y_t$ denote the x and y coordinates of worker w_i at time t . $d(w_i, w_j, t) = \sqrt{(w_i.x_t - w_j.x_t)^2 + (w_i.y_t - w_j.y_t)^2}$ denotes the Euclidean distance between w_i and w_j at time t on the communication network. If $d(w_i, w_j, t) \leq r_{prop}$ where r_{prop} is a pre-defined propagation radius, $w_i, w_j \in W$ are called neighbors on the communication network at time t .

The social network and the communication network can be collectively described as a hybrid network two-layer graph $G_{SC}^t = (W, E_f, E_n^t)$ by worker's ID. The friend relationship is constant, while the neighbor relationship changes over time. We assume that if w_i and w_j are friends on the social network or neighbors on the communication network, the influence can propagate from an influenced worker w_i to worker w_j .

An example of the network model is shown in Fig. 1. In the figure, the triangle indicates the location where the task is issued, the circle indicates the worker, and the shadowed region indicates the range of the task's influence radius during the initial propagation period. Once the task is issued, the task is pushed to the workers in the shadowed region. Worker 2 and Worker 1 are neighbors on the communication network, so the influence will be propagated from Worker 2 to Worker 1. Although Worker 3 is not in the influencing region, Worker 3 and Worker 2 are friends on the social network, so the influence will propagate to Worker 3.

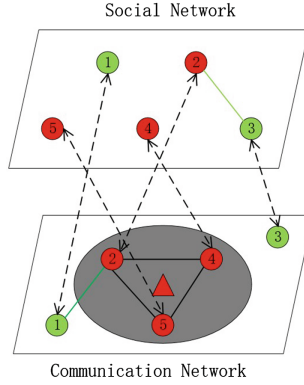


Fig. 1. The instance of the network model.

3.2 Problem Definition

Definition 3 (Task). The task can be represented by $T = \langle (T.x, T.y), T.content, T.radius, T.incentive, T.area, T.init.time, T.add.time \rangle$, respectively indicating the x and y coordinates of the task's issued location, the task's content, the task's influence radius, the reward for completing the task, the sensing area covered by the task, the issue time of the task, the task's initial propagation duration, and the task's additional propagation duration. Therefore, the initial propagation period of the task is $[T.time, T.time + T.init.time)$ and the additional propagation period of the task is $[T.time + T.init.time, T.time + T.init.time + T.add.time)$. We use $C(T.x, T.y, T.radius) = \{(x, y) | (x - T.x)^2 + (y - T.y)^2 \leq T.radius^2\}$ to denote the circle region around $(T.x, T.y)$ with radius $T.radius$.

Definition 4 (Worker). The worker can be represented by $W = \langle (w.ID, w.t, w.x_t, w.y_t), w.ability, w.reward \rangle$, respectively indicating a check-in record of the worker, the worker’s ability, the worker’s expected reward. We assume that each worker is honest, and the uploaded sensing data is credible. We don’t consider workers moving from outside the sensing area to the target sensing area.

Definition 5 (Spatial Coverage). The publisher specifies a set of subareas $F = \{f_1, f_2, \dots, f_m\}$. A subarea is considered to be covered when at least one sensing data is obtained.

Task Assignment Problem Definition. First, given a hybrid network two-layer graph $G_{SC}^t = (W, E_f, E_n^t)$, a sensing task T , workers’ check-in data, the publisher selects the seed nodes during the initial propagation period. Then, the workers in the seed set spread the task to their friends on the social network or their neighbors on the communication world during the additional propagation period. Task Assignment is to find the final recruited worker set $R(S)$ through the influence of the seed set S with the purpose of maximizing the spatial coverage, expressed as

$$\begin{aligned} & \text{Maximize } \frac{|Coverage(R(S))|}{|F|} \\ & \text{Subject to } |S| \leq \text{max_seed and } |R(S)| \leq \text{max_worker} \end{aligned}$$

where max_seed is the maximum number of seed set and max_worker is the maximum number of final recruited workers.

Theorem 1. *Task assignment problem proposed in this paper is an NP-hard.*

Proof. The problem of influence maximization on social networks has been proved to be an NP-hard. The problem of influence maximization on social networks can be reduced to the problem of spatial coverage maximization, which is the goal of task assignment. The problem of maximizing spatial coverage borrows the idea of influence propagation on social networks and communication networks in order to recruit workers to assign the task to cover as many subareas as possible. The process of finding workers is NP-hard. Therefore, task assignment problem proposed in this paper is an NP-hard.

Therefore, this paper proposes a greedy algorithm to solve the problem of MCS task assignment.

4 Solutions to Task Assignment Problem

4.1 Recruitment Probability Calculation

According to WeberFeckner’s law in the field of psychophysics [9], with the increase of external stimulus, people’s sensing is increased but the degree of enhancement decreases. In order to maintain this property, the influencing probability function used in this paper is defined as:

$$I(x, I_{\max}) = (I_{\max} - 1)\sqrt{1 - (1 - x)^2} + 1 \quad (1)$$

where x is the input probability increasing parameter and I_{\max} is the maximum increase. For $x \in (0, 1)$, we have $I(0, I_{\max}) = 1$, $I(1, I_{\max}) = I_{\max}$.

We consider several factors from the perspective of the task publisher:

Worker's Ability. $I_1(T, w) = I(x, I_{\max 1})$ is used to measure the impact of worker's ability, where $I_{\max 1}$ is the upper limit of increase. The more similar the task's content is to the worker's ability, the higher the probability that the publisher will recruit the worker. x is expressed as

$$x = Jaccard(T.content, w.ability) = \frac{|T.content \cap w.ability|}{|T.content \cup w.ability|} \quad (2)$$

Incentive Mechanism. $I_2(T, w) = I(x, I_{\max 2})$ is used to measure the impact of the incentive mechanism, where $I_{\max 2}$ is the upper limit of increase. The closer the worker's expected reward to the task's reward, the higher the probability that the publisher will recruit the worker. x can be defined as

$$x = Match(T, w) = \begin{cases} \frac{1}{1+e^{-(T.award-w.expect)}}, & T.award \geq w.expect \\ 0, & T.award < w.expect \end{cases} \quad (3)$$

Stay Time. $I_3(T, w) = I(x, I_{\max 3})$ is used to measure the impact of the worker's stay time, where $I_{\max 3}$ is the upper limit of increase. The longer the worker stays in the subarea, the more likely it is that the task will be completed and the probability that the publisher will recruit the worker. x can be expressed as

$$x = Stay(T, w) = \begin{cases} \frac{duration(T, w)}{T.init.time}, & \text{during the initial propagation period} \\ \frac{duration(T, w)}{T.add.time}, & \text{during the additional propagation period} \end{cases} \quad (4)$$

The Number of Influenced Times. $I_4(T, w) = I(x, I_{\max 4})$ is used to measure the impact of the number of worker's influenced times, where $I_{\max 4}$ is the upper limit of increase. The more times the worker is influenced, the higher the probability that the publisher will recruit the worker. x can be defined as

$$x = Frequency(T, w) = \begin{cases} 1, & \text{during the initial propagation period} \\ \min(\frac{n(T, w)}{n_{\max}}, 1), & \text{during the additional propagation period} \end{cases} \quad (5)$$

where $n(T, w)$ is the number of times worker w receives task T during the propagation period, and n_{\max} is the maximum number of times all workers are influenced during the propagation period.

Worker's Movement. $I_5(T, w) = I(\min(\frac{num}{num_{\max}}, 1), I_{\max 5})$ is used to measure the impact of worker's movement, where $I_{\max 5}$ is the upper limit of increase, num is the number of workers in the subarea that the worker will arrive at the next moment, and num_{\max} is the maximum number of workers in all subareas

at the next moment. If the worker moves to the subarea with sparse density at the next moment, the probability that the publisher will recruit the worker is higher.

Therefore, the probability that the publisher recruits workers to complete task T is

$$P(T, w) = \min(P_0 \times I_1(T, w) \times I_2(T, w) \times I_3(T, w) \times I_4(T, w) \times I_5(T, w), 1) \tag{6}$$

where P_0 is the fundamental influence probability. This paper sets a probability threshold $P_{threshold}$, and if $P(T, w) > P_{threshold}$, it is considered that worker w can be recruited by the publisher.

4.2 Offline Worker’s Movement Prediction Algorithm

Definition 6 (Heterogeneous Information Network). Heterogeneous information network can be represented by $G = (V, E)$, where V represents a set of nodes, E represents a set of edges. The network mode can be represented by $TG = (A, R)$, where A represents a set of node types, and R represents a set of edge types.

In the process of offline worker’s movement prediction, the network is modeled as a heterogeneous network with weights, as shown in Fig. 2, denoted by $G(W, L, E, W)$. W denotes the set of workers and L denotes the set of check-in locations. $E = E_{WL} \cup E_{LL}$ denotes the set of edges, where E_{WL} denotes the worker’s check-in behavior and E_{LL} denotes the correlation between check-in locations. $W = W_{WL} \cup W_{LL}$ denotes the set of edges’ weights.

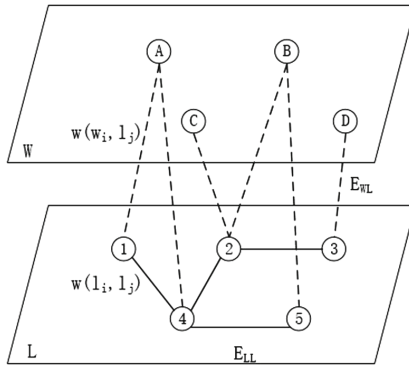


Fig. 2. The structure of the heterogeneous network.

Given a time threshold Δt , if there is a worker’s consecutive check-in location l_i and l_j , and the check-in time interval between them is less than Δt , then l_i and l_j are considered to be relevant [10]. The weight between the worker and the

check-in location is the total number of times the worker has checked in to this location. The weight between the check-in location and the check-in location is the total number of consecutive visits to the two check-in locations within the time threshold Δt for all workers.

Definition 7 (Meta-path). Meta-path can be defined as $P = A_1 R_1 \longrightarrow A_2 R_2 \longrightarrow \dots R_n \longrightarrow A_{n+1}$. If there is a real path $p = (v_1, v_2, \dots, v_{n+1})$, then path p is called an instance path of meta-path P , all such real paths are called the instance path set P' of the meta-path set P .

According to the small world phenomenon [11] and the three-degree influence theory [12] can infer that the relationship between the meta-paths with lengths greater than three is very weak. Therefore, the meta-path set is expressed as $P_S = \{P_1, P_2\}$, where P_1 denotes the meta-path of “ $W - L - L$ ” (workers may go to locations related to the location they have checked in) and P_2 denotes the meta-path of “ $W - L - L - L$ ” (workers may go to multiple locations related to the location they have checked in).

For a meta-path P , the eigenvalue is the sum of the degrees of association of all instance path sets P' , and the correlation formula of the correlation path $Correlation(P)$ of the meta-path P is

$$Correlation(P) = \sum_{p \in P'} cor(p) \quad (7)$$

where $cor(p)$ is the degree of association between the first node and the tail node of the instance path p .

For an instance path $p = (a_1, a_2, \dots, a_{n+1})$, this paper calculates the correlation degree based on the idea that each jump in the random walk is considered to be independent of each other. The calculation formula of $cor(p)$ is

$$cor(p) = \prod_{i=1}^n prob(a_i, a_{i+1}) \quad (8)$$

where $prob(a_i, a_{i+1})$ is the probability from the node a_i to the node a_{i+1} .

The calculation formula of $prob(a_i, a_{i+1})$ is

$$prob(a_i, a_{i+1}) = \frac{w(a_i, a_{i+1})}{\sum_{v \in N(a_i)} w(a_i, v)} \quad (9)$$

where $N(a_i)$ is a set of nodes that are connected to the node a_i and are consistent with the type of the node a_{i+1} .

According to the meta-path set $P_S = \{P_1, P_2\}$, the degree of association of any worker's check-in location pair (w, l) is calculated. A correlation degree vector $\alpha = \{Correlation(P_1), Correlation(P_2), 1\}$ is obtained, where the constant 1 is added only by using the logical distribution formula. Therefore, the probability that worker w goes to the location l is

$$\rho = \frac{1}{1 + e^{-\alpha \cdot \theta}} \quad (10)$$

where the vector θ is the corresponding weight obtained by the two types of meta-paths based on the training set using the supervised learning method. Finally, the most likely check-in location is selected as the worker’s movement prediction.

The pseudocode of the worker’s movement prediction algorithm based on meta-path is shown in Algorithm 1, where L_C represents check-in location l where worker w may arrive at the next moment.

Algorithm 1. Worker’s movement prediction algorithm

Input: The meta-path set $P_S = \{P_1, P_2\}$, Heterogeneous network $G(W, L, E, W)$, The target worker w

Output: Worker’s movement prediction location l

```

1:  $L_C = findLocation.Set(w)$ 
2: while  $l \in L_C$  do
3:    $i = 0$ 
4:   while  $P \in P_S$  do
5:      $Correlation_P = 0$ 
6:      $P' = findInstancePathSet(w, P)$ 
7:     while  $p \in P'$  do
8:        $cor_p = computeCorrelation(p)$ 
9:        $Correlation_p+ = cor_p$ 
10:    end while
11:     $\alpha[i+] = Correlation_p$ 
12:  end while
13:   $\alpha[i] = 1$ 
14:   $\rho_l = e^{\alpha\theta} / (e^{\alpha\theta} + 1)$ 
15: end while
16:  $l = top - one(\rho_l)$ 
17: return  $l$ 

```

4.3 Task Assignment Algorithm

Main Idea. This paper proposes a greedy task assignment algorithm named H-GTA, which first selects the seed nodes and then iteratively selects workers according to the seed nodes.

Algorithm Details. The main steps of the H-GTA algorithm are shown in Algorithm 2. In line 1, $W_{init} = \{w \in W | (w.x_t, w.y_t) \in C(T.x, T.y, T.radius)\}$ represents the set of workers within the region of the task’s influence radius during the initial propagation period. We can calculate the probability of recruiting workers according to Eq. 6 to find the candidate seed set $W_{initCandidate}$ that can be recruited as a worker. In line 2, we initialize the seed set to be empty. In lines 3–7, we select workers according to the heuristic function $Heuristic(w) = \alpha \times InfluenceDegree(w, W_{initCandidate}) + (1 - \alpha) \times (|Coverage(S \cup w)| -$

$|Coverage(S)|$), where $InfluenceDegree(w, W_{initCandidate})$ denotes the influence of worker w in the candidate seeds, $(|Coverage(S \cup w)| - |Coverage(S)|)$ denotes the utility of spatial coverage that add w into the seeds S , which α is a parameter used to balance the influence of the worker and the utility of spatial coverage. The selected worker is added to the seed set S , and the worker is removed from the candidate seed set $W_{initCandidate}$, and the next cycle is entered until max_seed workers are selected can end the loop. In line 8, the workers in the seed set S of the previous step propagate the task to their friends through the social network or neighbors through the communication network during the additional propagation period. Calculate the probability of recruiting workers according to Eq. 6 again to find the candidate workers set $W_{addCandidate}$ that can be finally recruited as workers. In line 9, workers in the seed set S are added to the final recruited workers set $R(S)$. In lines 10–18, $Utility(w) = |Coverage(R(S) \cup w)| - |Coverage(R(S))|$ denotes the utility of the spatial coverage that add w to the selected workers set $R(S)$. If the utility is zero, the loop is jumped out, otherwise the selected worker is added to the final recruited workers set $R(S)$ and deleted from the candidate workers set $W_{addCandidate}$, then enters the next cycle until $(max_worker - max_seed)$ workers are selected can finish the loop. In line 19, return the final recruited workers set $R(S)$.

Algorithm 2. H-GTA algorithm

Input: The workers set W_{init} , The maximum number of seeds max_seed , The maximum number of recruited workers max_worker

Output: The selected workers set $R(S)$

```

1:  $W_{initCandidate} = findCandidate(W_{init})$ 
2: set  $S = \emptyset$ 
3: while  $|S| \leq max\_seed$  do
4:   select  $w$  from  $W_{initCandidate}$  with maximum  $Heuristic(w)$ 
5:    $S = S \cup \{w\}$ 
6:    $W_{initCandidate} = W_{initCandidate} - \{w\}$ 
7: end while
8:  $W_{addCandidate} = findCandidate(networkPropagate(S))$ 
9:  $R(S) = S$ 
10: while  $|R(S)| \leq max\_worker$  do
11:   select  $w$  from  $W_{addCandidate}$ 
12:   if  $Utility(w) == 0$  then
13:     break
14:   else
15:      $R(S) = R(S) \cup \{w\}$ 
16:      $W_{addCandidate} = W_{addCandidate} - \{w\}$ 
17:   end if
18: end while
19: return  $R(S)$ 

```

5 Experimental Study

5.1 Experimental Setup

Datasets. This paper selects the real LBSN datasets Brightkite and Gowalla because they include friend relationship between workers and contain worker’s movement. Brightkite contains 2,627,870 check-in records made by 58,228 users involving 772,933 locations. Among all the users there are in total 214,078 friendship links. Gowalla contains 6,264,203 check-in records made by 196,591 users involving 1,280,956 locations. The dataset also contains 950,327 friendship links among users [13]. Due to the small moving distance of workers and the large scale of the communication world, it is difficult to observe the movements of workers. To accurately analyze workers’ movements, we use a portion of the original Brightkite and Gowalla. The relevant details of the datasets are shown in Table 1.

Table 1. Datasets.

	Brightkite	Gowalla
Number of workers	8,650	10,693
Number of edges	32,536	55,506
Number of check-in records	458,648	333,915
Number of average check-in records	53.023	31.227

According to the datasets, the distribution of distances between different check-in locations of workers can be measured. As shown in Fig. 3, we can find that the location of workers on the communication network remains unchanged in many cases. Therefore, workers’ movements after they are recruited will not affect the completion of the task.

According to the datasets, the number of neighbors on the communication network within different distances can be measured. As shown in Fig. 4, it can be seen that there are many neighbors near a worker on the communication network.

Sensing Area. The sensing area covered by the given task is a 240 km \times 200 km rectangle region, as shown in Fig. 5. Assumed that the task is published at longitude -75.2 and latitude 40 , which is taken as a two-dimensional coordinate (120, 100) point. It is assumed that the whole sensing area is divided into 480 subareas equally, each subarea is 10 km \times 10 km, and the subareas in the ocean are removed. Finally, 344 virtual subareas are considered in this paper.

The relevant details of the parameter settings are shown in Table 2.

Baselines. The following baseline methods are selected for comparative experiments:

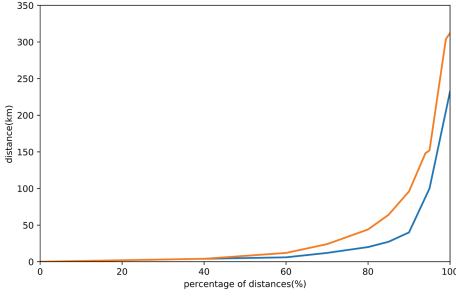


Fig. 3. The distribution of distances.

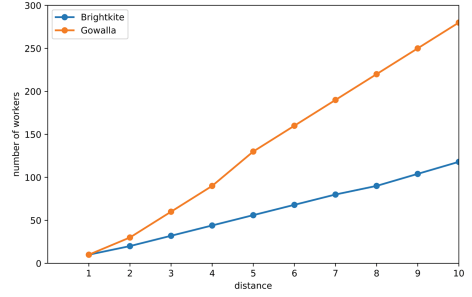


Fig. 4. The distribution of number of neighbors.

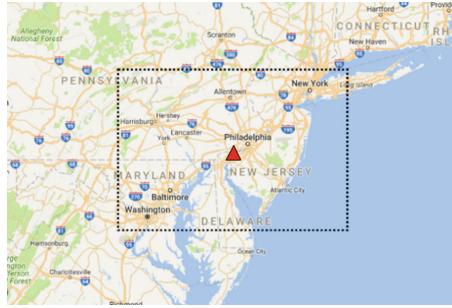


Fig. 5. The whole sensing area.

- *MaxDegree*: This method iteratively selects workers who have the maximum degree on the social network.
- *MaxCov* [14]: This method iteratively selects workers with the goal of maximizing the spatial coverage.
- *Heuristic Greedy (HG)*: This method adopts a greedy algorithm to iteratively select workers with the maximum heuristic function.
- *NaiveFast*: This method uses a greedy algorithm based on the intuition-based rank utility function to select seed nodes.
- *Basic-Selector* [2]: This method selects the most beneficial seeds iteratively according to predictive temporal-spatial coverage in a greedy way.
- *Fast-Selector* [2]: This method uses a two-stage fast seed selection algorithm to iteratively select workers.

Implementation. All the algorithms are implemented in Python.

5.2 Experiment Results

In this paper, spatial coverage and running time are used to evaluate the performance of the algorithm.

Table 2. Parameter settings.

Parameters	Settings
r_{prop}	10
$T.radius$	10
P_0	Randomly generate from [0.1, 0.5]
I_{max}	$I_{max1} = 3, I_{max2} = 1.5, I_{max3} = 1.5, I_{max4} = 6, I_{max5} = 3$
α	Brightkite:0.64, Gowalla:0.56
max_seed	25, 50, 75, 100
max_worker	500

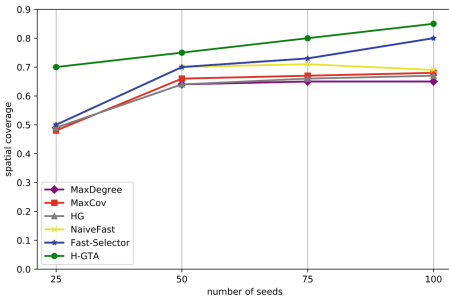


Fig. 6. Brightkite’s spatial coverage.

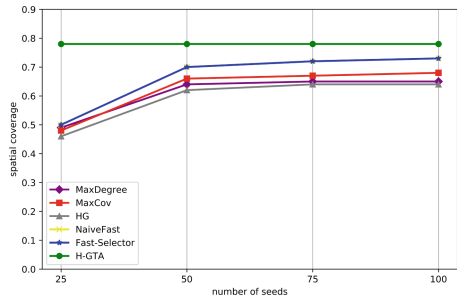


Fig. 7. Gowalla’s spatial coverage.

Figures 6 and 7 compare spatial coverage between the H-GTA algorithm and other comparison algorithms under different datasets and the different number of seeds settings.

As shown in Figs. 6 and 7, the H-GTA algorithm proposed in this paper can obtain higher spatial coverage than other comparison algorithms and have little to do with the number of seeds.

Figures 8 and 9 compare the spatial coverage of the H-GTA algorithm with or without movement prediction.

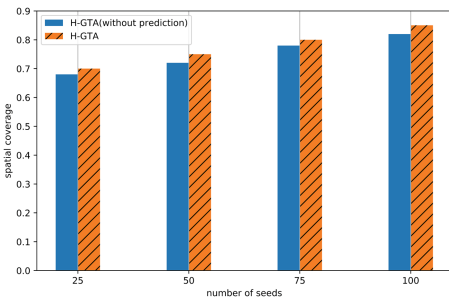


Fig. 8. Brightkite’s spatial coverage with or without movement prediction.

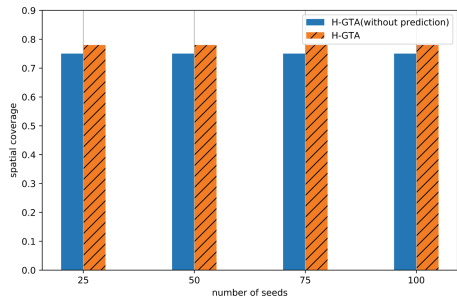


Fig. 9. Gowalla’s spatial coverage with or without movement prediction.

It can be seen from Figs. 8 and 9 that the H-GTA algorithm with movement prediction obtains slightly higher spatial coverage than the algorithm without movement prediction.

Figures 10 and 11 compare running time between the H-GTA algorithm with other comparison algorithms.

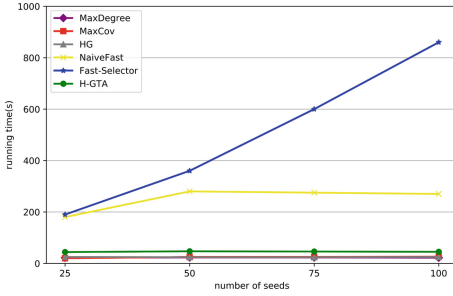


Fig. 10. Brightkite's running time.

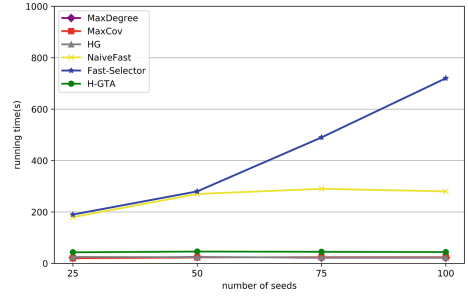


Fig. 11. Gowalla's running time.

It can be seen from Figs. 10 and 11 that the H-GTA algorithm proposed in this paper can obtain higher spatial coverage than other comparison algorithms in the case of not running for a long time.

6 Conclusion

This paper proposes a solution to the problem of task assignment for environmental monitoring MCS tasks. This method draws on the idea of social influence to recruit workers to assign the task by social networks and communication networks, with the goal of maximizing the tasks spatial coverage. The experimental results verify the performance of the H-GTA algorithm.

References

1. Wang, J., Wang, L., Wang, Y., Zhang, D., Kong, L.: Task allocation in mobile crowd sensing: state of the art and future opportunities. *IEEE Internet Things J.* **5**(5), 3747–3757 (2018)
2. Wang, J., Wang, F., Wang, Y., Wang, L., Qiu, Z.: Social-network-assisted worker recruitment in mobile crowd sensing. *IEEE Trans. Mob. Comput.* **18**(7), 1661–1673 (2019)
3. Zhang, M., et al.: Quality-aware sensing coverage in budget constrained mobile crowdsensing networks. *IEEE Trans. Technol.* **65**(9), 7698–7707 (2016)
4. Karaliopoulos, M., Telelis, O., Koutsopoulos, I.: User recruitment for mobile crowdsensing over opportunistic networks. In: *INFOCOM*, pp. 2254–2262 (2015)
5. Xiong, H., Zhang, D., Wang, L., Chaouchi, H.: EMC3: energy-efficient data transfer in mobile crowdsensing under full coverage constraint. *IEEE Trans. Mob. Comput.* **14**(7), 1355–1368 (2015)

6. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 26–29, San Francisco, CA, USA (2001)
7. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. *Theory Comput.* **11**, 105–147 (2015)
8. Li, J., Cai, Z., Yan, M., Li, Y.: Using crowdsourced data in location-based social networks to explore influence maximization. In: INFOCOM, pp. 1–9, San Francisco, CA, USA (2016)
9. Dehaene, S.: The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends Cogn. Sci.* **7**(4), 145–147 (2003)
10. Cao, J., Dong, Y., Yang, P., Zhou, T., Liu, B.: POI recommendation based on meta-path in LBSN. *Chin. J. Comput.* **39**(4), 675–684 (2016)
11. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. *Nature* **393**(6684), 440–441 (1998)
12. Susan, K.W.: Connected: the surprising power of our social networks and how they shape our lives. *Chin. J. Comput.* **3**(3), 220–224 (2011)
13. Wang, H., Terrovitis, M., Mamoulis, N.: Location recommendation in location-based social networks using user check-in data. In: 21st SIGSPATIAL International Conference on Advances in Geographic Information Systems, Orlando, FL, USA, pp. 364–373 (2013)
14. Zhang, D., Xiong, H., Wang, L., Chen, G.: CrowdRecruiter: selecting workers for Piggyback crowdsensing under probabilistic coverage constraint. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing, Seattle, WA, USA, pp. 703–714 (2014)



A Barrier Coverage Enhancement Algorithm in 3D Environment

Xiaochao Dang^{1,2}, Yuexia Li¹, Zhanjun Hao^{1,2(✉)}, and Tong Zhang¹

¹ College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, Gansu, China

zhanjunhao@126.com

² Gansu Province Internet of Things Engineering Research Center, Northwest Normal University, Lanzhou 730070, Gansu, China

Abstract. Barrier gap repair in wireless sensor networks is an inevitable problem in barrier coverage research. This paper studies the repair of the barrier gap of wireless sensor networks in 3D environment and improves the coverage performance. This paper proposes a distributed barrier gap repair method in 3D environment. Firstly, look for the repair chain in a three-dimensional environment through an improved ant colony algorithm. Then, construct a relationship model between the mobile node and the patch chain, and then select the optimal barrier gap repair location. Finally, the priority model of gap filling is set by the moving distance of the node and whether it is the key area, and the corresponding repair of the gap is performed. The results of real experiments and simulation experiments show that the proposed algorithm can reduce the number of nodes used and reduce the mobile energy consumption when the mobile node repairs the gap.

Keywords: Three-dimensional barrier · Gap repair · Repair chain · Improved ant colony algorithm · Partition

1 Introduction

In the wireless sensor network, due to the limited energy of the node itself and the influence of the environment, the barrier coverage may generate a barrier gap after working for a period of time. Barrier coverage is mainly used for the monitoring of intrusive targets [1]. The existence of the barrier gap will cause the barrier coverage to reduce the performance of event monitoring. How to repair the gap between barriers has always been a hot topic of research. With the development of technology, moving nodes to locations to be repaired is no longer a problem. For example, a flying robot developed can fly to a fixed position and perform corresponding work. This makes the mobile node to the repair location and then repair the gap called the important method of repairing the gap [2]. At present, there are two problems in the study of network construction cost and network lifetime [3]. The cost of the barrier clearance is not considered, and these are the study of the barrier gap in two-dimensional environment. In the three-dimensional environment, how to find the repair position of the barrier gap and how to repair the gap in time need to be solved. The existing research on the barrier

gap in the three-dimensional environment only repairs the barrier gap from the ideal environment [4], and does not consider the problem of deployment node waste caused by the slope in the three-dimensional environment. On the basis of previous studies, this paper proposes a distributed barrier coverage gaps repairing algorithm (DBCR) for three-dimensional environment on the basis of ensuring that the barrier gap can be completely repaired. Two-dimensional meshing, finding the gap of the barrier through the probability perception of the node, constructing the barrier gap repair chain through the improved ACO algorithm, and constructing the relationship model between the node and the position to be filled according to the distance and slope problem, and using the relationship model to determine the optimal. The barrier is patched and repaired accordingly. In the three-dimensional environment, the energy-saving and efficient moving nodes are used to realize the repair of the barrier gap. The contributions of this paper are:

- (1) According to the improved ACO algorithm, a repair path of the barrier gap is created, and a repair chain of the barrier gap is constructed.
- (2) A relationship model between the mobile node and the gap to be repaired is proposed. Through this relationship model, the mobile node selects the repair location with the smallest distance in the repair chain.
- (3) According to the moving distance and whether it is the key area, create a priority model of the barrier gap to achieve efficient repair of the barrier gap in the three-dimensional environment.

Section 1 of this paper introduces related work, and Sect. 2 introduces nodes and network models. Section 3 details the improved barrier gap patching algorithm. Section 4 evaluates the performance of the proposed algorithm through simulation experiments. Section 5 summarizes the full text and introduces the next step.

2 Related Work

The application of barrier coverage in wireless sensor networks is becoming more and more extensive, such as national border monitoring, security surveillance and intrusion detection, with the aim of detecting intruders trying to cross protected areas [5]. In border monitoring, barrier coverage can be used to monitor enemy incursions; In forestry protection, barrier coverage can be used to monitor the spread of fire; In water regulation, barrier coverage can be used to monitor invasion of alien species; In industrial production, barrier coverage can be used to monitor the leakage of industrial materials [6]. Barrier coverage is mainly used for intrusion detection. When there is a barrier gap in the barrier coverage, the resources in the wireless sensor network cannot be fully utilized, which may lead to network loss of event monitoring [7]; In order to ensure the accuracy of intrusion monitoring, the sensor network must ensure the integrity of the entire network, then the patch clearance is an inevitable problem.

In recent years, in the study of barrier gap repair, the use of mobile sensor nodes to repair the barrier gap, and then improve the barrier coverage performance, has attracted extensive attention. Reference [8] proposes a barrier covering method for monitoring from external invasion and internal target breakthrough, called target barrier coverage;

Reference [9] repair the gap between barriers by adjusting the perceptual direction of nodes; The document [10] selects the node with the most neighbor to repair the barrier according to the relevant position information of the neighbor node; In [11], a directed strong K-barrier overlay padding algorithm (DSBCSB) based on the selection of directional nodes is proposed. The target node location is used as a reference to fill the directional node selection box for patch repair. In [12], the coverage holes of arbitrary line segments are studied, and the shortest k-covered line segments and the longest k-uncovered line segments are studied. The literature [13] designed two rotation algorithms to fix the gaps and repair the gaps. In [14], the gap repair problem in hybrid sensor networks is studied, and the coverage gap is repaired by the minimum-maximum scheme and the maximum lifetime scheme. In reference [15], a method of generating scheduling set is proposed by using the information redundancy caused by random deployment nodes, which realizes the repair of barrier coverage by removable nodes; In reference [16], a coverage gap repair algorithm (CGR), is proposed, which detects the low energy nodes in the network and sends the nearest movable nodes around the low energy nodes to replace them. The above-mentioned research, all is the barrier gap repair in the two-dimensional environment.

In the patch clearance problem, it is mainly considered from the aspects of reducing the mobile node and reducing the moving distance. For example, In [17], the directed sensor is modeled on the static sensor in the network, and the distance from the mobile node to the repair position is calculated, and then the mobile sensor with the shortest moving distance is selected for the gap repair. Literature [18] proposed a greedy mobile algorithm for heterogeneous wireless sensor networks, effectively scheduling different types of mobile sensors to different gaps while minimizing the total cost of movement. In [19], an optimization method for barrier clearance repair is proposed. The actual node topology is transformed into the demand topology of the number of mobile nodes, and the KSP algorithm is used to calculate the minimum number of mobile nodes to repair the obstacle gap. So far, few people have studied the repair of the gap in the three-dimensional environment.

In the past, the study of barrier coverage was mainly based on the two-dimensional environment, considering only the distance from the mobile node to the moving position, and was not suitable for the repair of the barrier gap in the three-dimensional environment. On the other hand, most of the research mainly focuses on the mobile node considerations to repair the barrier gap, and does not consider the route of the intrusion object invasion. The intrusion object is often chosen to be more concealed and traversed with respect to the path with shorter path. The barrier coverage area is divided into key areas, sub-key areas, etc. When the barrier gap appears, the gap is filled before the key monitoring area. For the gap repair, the ant colony algorithm can be used to find the optimal path and then the node is scheduled to repair the network [18]. However, this algorithm does not consider the waste of deployment nodes caused by the gradient in the three-dimensional environment, and is a centralized algorithm that requires topology information of the entire network and is not suitable for large networks in a three-dimensional environment. Based on this research, this paper proposes a distributed barrier gap repair algorithm suitable for three-dimensional environment: 2D meshing of 3D environment, construction of barrier gap repair chain by improved ACO algorithm, and according to distance and slope problem Construct a relationship

model between the node and the location to be filled, and use the relationship model to determine the optimal barrier repair location, the mobile node fills the barrier gap, and achieves the minimum node and minimum energy consumption to repair the three-dimensional barrier gap.

3 Omnidirectional Strong Barrier Coverage Model

Based Here, we assume that the deployment area is A , there are N_s static nodes and N_m dynamic nodes in A , and the nodes can obtain their own location information $p_i(x_i, y_i, z_i)$ in the network. Among them, redundant nodes are sufficient, and since the mobile node fills the movement of the barrier gap, no new gap is generated. A schematic representation of a barrier covering in a three dimensional environment is given herein (see Fig. 1). Among them, the gray circle represents the working node on the original barrier, the black circle represents the dynamic node that can move the gap, and the dotted line represents the barrier gap. It is assumed here that the sensor nodes in the WSN are isomorphic nodes, that is, the initial energy of each node, and the energy consumption and perceptual power per unit of motion are the same.

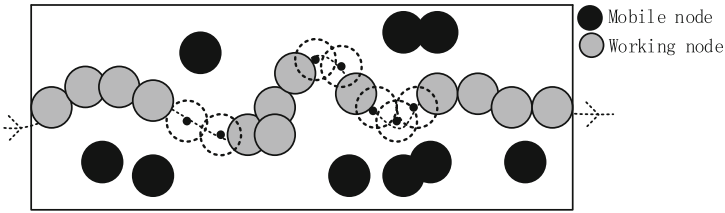


Fig. 1. Schematic diagram of the strong 1-barrier gap

3.1 Node-Aware Model

In general, the sensor node's perceptual ability changes mainly as the distance changes. Therefore, for the sensor node perception model, this paper uses the omnidirectional probability node perception model.

Definition 1 (node probability perception model). The perceptual model of the node is a probabilistic perceptual model, that is, the probability that the node perceives the intruder decreases as the distance between the intruder and the node increases. As shown in any point q within the node sensing range, node p can perceive point q Probability P_q is:

$$P(s_i, t) = \begin{cases} 1 & r > d(s_i, t) \\ e^{-\lambda\alpha^\beta} & r \leq d(s_i, t) \leq R_s \\ 0 & R_s < d(s_i, t) \end{cases} \quad (1)$$

$\alpha = d(s_i, t) - r$, λ and β are the sensory parameters of the sensor. R_s is the maximum perceived radius of the node, r is the radius that the node can perceive 100%, and $d(s_i, t)$ is the distance from the target event q to the node.

3.2 Network Patching Model

If the barrier gap is not repaired in time, some monitoring events may be missed. If such problems occur in military applications, the losses will be unavoidable. In this paper, a distributed barrier coverage gap repair algorithm is proposed for the three-dimensional environment. The algorithm firstly meshes the 3D environment in two dimensions, and finds the optimal repair chain of the barrier through the improved ACO algorithm. Then, the optimal repair location is found through the relationship model between the mobile node and the patching link. Finally, priority filling is based on whether the moving distance and the gap are key areas. The diagram of the repair of the fence gap is shown in the figure (see Fig. 2):

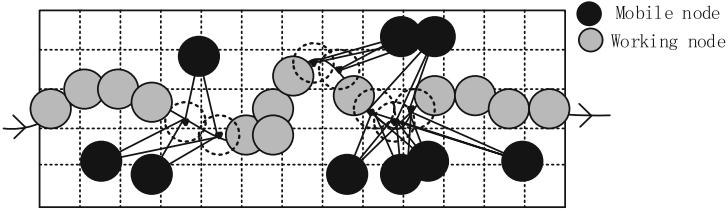


Fig. 2. Schematic diagram of barrier gap repair

In order to make better progress in the follow-up work, we give the following relevant definitions:

Definition 2 (Patch Chain). The connection of each hop nodes on the optimal paths sought by the improved ant colony algorithm IPACO is called the patch chain.

Definition 3 (displacement weight). The linear distance between the current node and the position of the gap to be filled, called the displacement weight.

Definition 4 (node utilization). When the patch gap is filled completely, the repair performance is measured by the number of mobile nodes. The more mobile nodes, the worse the node utilization of the network, because we aim to repair the barrier gap with the fewest number of nodes to achieve the least amount of nodes and achieve the lowest energy consumption. Therefore, there are:

$$C = \frac{n_m}{N_m} \tag{2}$$

Among them, the mobile node that fills all the gaps of the fence is n_m , N_m filling all the gaps of the barrier are all mobile nodes.

Definition 5 (Patch Chain Domain). Find the optimal path based on IPACO, the patch chain. We draw a circle with a radius r on the patch chain, and as long as a node moves onto the circle, the circle can be patched. Therefore, all the dotted circles have a moving node moving to all the circles of the patch chain, so the patch chain domain can be completely repaired. The dotted circle area is the repair chain domain (see Fig. 2).

In the three-dimensional environment, if the barrier coverage is more than the gap, the random moving node for gap repair may cause a lot of unnecessary energy waste. The traditional ant colony algorithm involves few parameters, is easy to implement, and more importantly, has the characteristics of finding the optimal path in complex path planning. So this paper uses the improved ant colony algorithm to plan the optimal path, and then build the repair chain. The mesh gradient model is introduced here first, and the patch chain is introduced later. When patching a 1-strong barrier in a dynamic network WSN, we divide the divided deployment area into several grids, and the representation of the steepness of each grid on the two-dimensional environment on the two-dimensional environment is called a grid gradient. Then there is the following formula:

$$i = \frac{h}{l} = \tan\vartheta \quad (3)$$

$$g = \left. \frac{\partial h(x,y)}{\partial x} \right|_{x_0} \cos\psi + \left. \frac{\partial h(x,y)}{\partial y} \right|_{y_0} \cos\left(\frac{\pi}{2} - \psi\right) \quad (4)$$

Where i is the slope, h is the vertical height of the slope, and l is the horizontal distance. ϑ is the angle between the slope and the horizontal plane. g is the grid gradient of the grid slope of the sensor node p along the ψ direction, $h(x,y)$ is the elevation at the coordinate (x,y) , and ψ is the angle with the slope direction. When $\psi = 0$ is, the gradient and the slope are equal. After the target area is divided into several grids, the grid gradient in which each node is located is obtained and placed in the grid gradient set G , and then the spatial weighting factor is introduced. The spatial weighting factor is described in detail later.

4 DBCR

In this paper, a distributed barrier coverage gap repair algorithm is proposed for the three-dimensional environment. The algorithm mainly includes the improved ant colony algorithm to find the repair chain, the structure of the barrier gap repair chain domain, the construction of the mobile node and the repair chain relationship model, and the repair of the barrier gap. So this chapter starts from how to improve the ACO from the grid gradient and build the relationship model between the mobile node and the barrier repair chain. It introduces how to improve the ant colony algorithm to find the repair chain, how to construct the mobile node and the barrier repair chain. The relationship model is used to determine the location of the patch; then, the performance of the proposed algorithm is introduced from the gap and correlation analysis of the gap and whether the gap is the key area.

4.1 Improve the ACO Algorithm

We have to repair the gap between the barrier, we must first find the optimal path. This paper improves the traditional ant colony algorithm by introducing the spatial weighting factor and limiting the ant mobility ability, so as to find the repair chain.

Definition 6 (space weight). With the starting node as the horizontal plane, the grid on the path to be selected is given different weights, called the spatial weight.

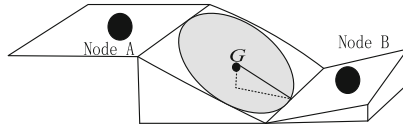


Fig. 3. Based on the spatial weight node selection graph

The optimization steps of the spatial weights are given below.

(1) Space weight

This paper uses the spatial weighting factor introduced by the grid direction gradient and uses it as another inspirational information of IPACO. Firstly, the target area is divided into several grids; from Sect. 1, according to the position information of the redundant nodes and the points to be filled, the corresponding grid gradient can be obtained; finally, the corresponding space weights are obtained. Here, we use ζ_{ij} to represent the spatial weight. Then there are:

$$\zeta_{ij} = \frac{G_{ij} - g_{ij}}{g_{ij}} \tag{5}$$

G_{ij} represent the grid gradient of the current grid, and g_{ij} represents the grid gradient of the next hop node (see Fig. 3). when the distance between the node A and the node B is equal to the position to be filled, if there is $g_A > g_B$ in the grid gradient of the node A and the node B, then $\zeta_A < \zeta_B$, that is, when the node A and the node B are selected to fill When the node of the gap position is selected, the node B with a gentle slope will be selected. That is to say, when the ant finds the optimal path, its selection probability will be affected by the spatial weight, that is, the smaller the grid gradient value g_{ij} , the larger the value of ζ_{ij} , the greater the state transition probability of the ant.

(2) Ant’s mobility

If there is a gap between the nodes, the concept of a strong barrier is not met. In order to construct a strong barrier, there must be overlapping parts of the adjacent nodes. Therefore, in order to ensure that the constructed barrier is a strong barrier, we limit the ant’s ability to $2R_r$. When there is no suitable node, the corresponding padding node position is selected and the next hop node is selected. According to (5), the state transition probability is changed as follows:

$$p_{ij}^k(t) = \begin{cases} \frac{\tau_{ij}^\alpha(t)\eta_{ij}^\beta(t)\zeta_{ij}(t)}{\sum_{s \in allowed_k} \tau_{is}^\alpha(t)\eta_{is}^\beta(t)\zeta_{is}(t)} & j \in allowed_k \\ 0 & j \notin allowed_k \end{cases} \quad (6)$$

The above formula (6) represents that the ant selects the state transition probability of the next node. The probability of transition probability $p_{ij}^k(t)$ of ant k moving from city i to node j at time t is the same as the meaning of other symbols in the formula, and will not be described here. IPACO starts from the left node and traverses the node. When there is no node in the next hop, each grid center point is selected as the virtual node, and the traversal is continued until the end point, the optimal path is recorded, and the repair chain is constructed.

In the traditional algorithm, the ant chooses the next node to rely on only two factors, τ_{ij} and η_{ij} . The positions of the nodes are different, and the heights may be different. After the target area is meshed, the different spatial weights of the nodes are obtained according to the grid gradient of the nodes, and the improved ant colony algorithm is used to make the ants search for the most from the left boundary. Excellent path, when encountering different paths can go, find the optimal repair chain by executing the improved ant colony algorithm. The pseudo code is as follows (Table 1):

Table 1. IPACO looking for patch chain pseudo code.

Input: Sensor node position coordinates in the deployment area
 Output: minimum barrier length , node position information pi

1. Data initialization
2. for $i=0$ to N do
3. for $j=0$ to N do
4. The ant searches for the barrier, and the barrier increases accordingly. After there is no optional node, the ant stops searching.
5. if $barrier > old_barrier$
6. $old_barrier = barrier$
7. end if
8. if $barrier = old_barrier$
9. if $length < old_length$
10. $old_barrier = barrier$
11. end if
12. end if
13. end if

4.2 Repair Chain Domain of the Barrier Gap

The fence gap repair is different from the area coverage void repair. In the fence gap repair, the fence gap repair is realized as long as the repair gap forms a strong fence. To achieve the repair of the fence gap, it is first necessary to determine the repair position of the fence gap. We will draw a circle with a radius r on the repair chain, called the patch circle, and the patch circle forms the patch chain domain. However, if the mobile node only moves to the patch chain location for repair, it may cause unnecessary energy loss. This paper mainly starts from reducing the moving distance of mobile nodes and improving node utilization. To facilitate the verification of the accuracy of the algorithm, the following theorem is given here:

Theorem 1: The node moves to a circle of radius r to repair the gap circle.

Proof: When the ant finds the repair chain and records the position information of each hop node of the ant, draws a circle with a radius of r . Because IPACO has limited the ant's range of activity to $2r$. Because the ant's range of motion is twice that of the patched circle, because $R = 2r$, that is, the radius of the moving node is twice that of the patching circle, the fence gap can be repaired when the node moves to the ant recording position. As shown in Fig. 4, the position on the repair chain that the ant looks for is the center of the circle, and the radius r is a circle. When the node moves to the circle, the circle can be repaired.

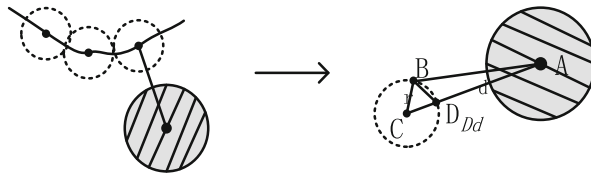


Fig. 4. Patch location

Theorem 2: When the moving node and the line determined by the center of the circle and the intersection of the circle drawn are the target position of the moving node, the mobile node can repair the gap of the barrier at a minimum cost.

Proof: A is the mobile node (see Fig. 4), the dotted circle C is the circle with the radius r on the repair chain, and the node A intersects the circle C on the repair chain at point D . Now take a point on the circle C , for B , make A, B, C three points, which is $\overline{AC} < \overline{AB} + \overline{BC}$ a triangle BCD . In any of the triangles, the sum of the two sides is greater than the third side, so that is, equivalent to $\overline{AD} + \overline{DC} < \overline{AB} + \overline{BC}$, therefore, there is $\overline{AD} + \overline{DC} < \overline{AB} + \overline{CD}$. The certificate is completed $\overline{AD} < \overline{AB}$.

In the past, research on the repair of barrier gaps did not consider the route of invasion of invading objects. Intruding objects tend to choose to be more concealed and traversed with respect to paths with shorter paths. So this paper introduces the concept of displacement weights. If you only rely on the moving distance to repair the gap, it may result in data loss caused by the key area not being repaired in time. This paper

considers the concept of displacement weight and division area. The barrier coverage area is divided into key areas, sub-key areas, etc. When the barrier gap appears, the gap is filled before the key monitoring area. When the barrier gap is repaired, the mobile node repair gap is performed according to the displacement weight and the regional level division, and the gap priority repair concept is proposed to repair the barrier gap. After the mobile node receives the information of the barrier gap, the distance between the repair chain and the repair chain is calculated first, which is the moving distance required by the mobile node to repair the gap, which is the distance that the mobile node needs to move to move to the target position, as shown in formula (8). As shown, the target location of the mobile node is determined. The detailed description of the repair location and related distance (see Fig. 4), where $d = AD$, $r = CD$, $Dd = AC$.

$$\chi_{ij}(t) = \frac{1}{d_{ij}} \quad (7)$$

$$d = Dd - r \quad (8)$$

$$\lambda_{si}^i = \frac{1}{d_{si}^i} + o(z) + \chi_{ij} \quad (9)$$

χ_{ij} represents the displacement weight. The priority of the barrier gap cluster is determined by Eq. (9). The closer the distance, the higher the level of the area and the higher the priority. Each mobile node generates a priority algorithm for each gap to select the highest priority gap patch for each mobile node.

4.3 Algorithm Performance Analysis

When searching for the optimal path in a relatively complex three-dimensional environment, the IPACO algorithm is used to find the repair chain, and then the relationship model between the mobile node and the repair chain is constructed. Finally, the priority gap is used to perform the barrier gap repair. In order to verify the superiority of the performance of the proposed algorithm, the performance of the proposed algorithm is analyzed by the improved convergence speed of the ant colony algorithm and the moving distance of the mobile node filling the gap of the barrier.

Analysis of IPACO Convergence Speed Problem. The traditional ant colony algorithm has a slow convergence rate due to lack of information in the early stage. Moreover, the traditional ant colony algorithm tends to select the nearest node when selecting the next hop node, and it is easy to fall into the local optimal solution. The IPACO algorithm introduces the spatial weight and the displacement weight by meshing, selects the node with the highest transition probability, and eliminates the node that is not applicable to the next hop node in advance, which reduces the number of iterations and improves the convergence speed. It can be seen that the IPACO algorithm proposed in this paper is more excellent in the three-dimensional environment.

Analysis of Moving Distance and Energy Consumption. Considering the difference between the actual 3D environment and the ideal environment, we find the shortest patching path by improving the ant colony algorithm, and then construct the relationship between the mobile node and the patching location to minimize the moving distance of the node while minimizing energy consumption. Thus, there are the following propositions:

Proposition: The algorithm in this paper can repair the barrier gap, but also achieve the minimum number of nodes and the smallest energy consumption.

Proof: Obviously, because the moving distance of the node is related to the repair chain, the repair chain is related to the ant's search path, that is, the ant's transition probability $P_{ij}^k(t)$ in the ant colony algorithm. We start from improving the transfer probability of ants. As shown in Eq. (5), when the spatial weights of the next hop nodes are different, and the other conditions are the same, the ants choose the probability that the next hop node tends to be straight. There are $p_{1ij}^k(t) < p_{ij}^k(t)$, the node is selected by Eq. (6). For the same reason, the algorithm constructed by this paper tends to be more straight and gentle barrier coverage. The patch chain can be reached to the shortest.

From Theorem 1, we can know that the movement of the mobile node when the mobile node reaches the target position is also the smallest. In the process of repairing the barrier gap, since the energy consumption of the redundant node mobile node is positively correlated with the distance moved by the node, the energy consumption of the node movement increases as the distance increases. The certificate is completed.

Compared with other barrier clearance algorithms, the main contributions of the proposed algorithm are as follows:

- (1) According to the improved ACO algorithm, a repair path of the barrier gap is created, and a repair chain of the barrier gap is constructed.
- (2) A relationship model between the mobile node and the gap to be repaired is proposed. Through this relationship model, the mobile node selects the repair location with the smallest distance in the repair chain.
- (3) According to the moving distance and whether it is the key area, create a priority model of the barrier gap to achieve efficient repair of the barrier gap in the three-dimensional environment.

5 Simulation Experiment

In order to verify the performance of the proposed method in the patch gap repair, the simulation experiment deployment area is a rectangular target area with a mesh division ratio of 10:1. There is already a barrier coverage with gaps in the area, and a number of redundant mobile nodes, where. The perceived probability of the node is $p = 7.5$. Here, a simulation experiment of 1-strong barrier gap repair is performed on the target area. Unless otherwise specified, the experimental parameter values are set in Table 2. The results of the following simulation experiments are taken as 100 random test averages, and the number of iterations is set to 300.

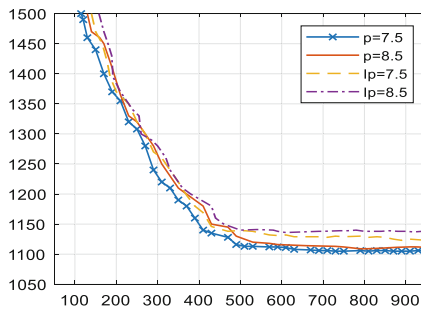
Table 2. Experimental parameters.

Experimental parameters	Ranges
Monitoring area A	1000 m \times 100 m
Pheromone concentration α	1.0
Volatilization coefficient β	1.0
Inspirational information ε	2.0
Total amount of information Q	100
Perceptual radius r	10 m
Degree of information attenuation ρ	0.7

Enter the grid gradients with different heights of each grid and start to find the optimal path. The node is then moved to the optimal location and the relevant data information is recorded. In order to better verify the superiority of the DBCR algorithm proposed in this paper, we select the traditional ant colony algorithm and the greedy algorithm widely used for barrier gap repair proposed in the literature [20]. Respectively, as ACO and Greedy, and The DBCR algorithm proposed in this paper is experimentally compared.

5.1 Algorithm Patching Performance

First, we compare the number of iterations that need to be repeated when the traditional ACO and IPACO algorithms converge. The simulation experiment was carried out under the same number of sensors (see Fig. 5). We hope that during the gap filling process, we choose a node that is more gradual and closer to the barrier gap filling. Therefore, IPACO preferentially selects redundant nodes with steeper slopes to fill. However, IPACO can reduce the number of iterations and improve the efficiency of the ant colony algorithm to repair the barrier gap. Nodes with different perceptual probabilities also have a certain impact on the IPACO algorithm to fill the gap of the barrier. The comparison of the iterations of ACO and IPACO under the perceived probability of different nodes (see Fig. 5).

**Fig. 5.** Comparison of IPACO and ACO iterations

5.2 Comparison with Other Algorithms

The DBCR algorithm mobile node number and node moving distance simulation experiment graph under the default parameters can increase the number of mobile nodes as the number of grids increases (see Figs. 6 and 7). This is because the traditional ACO cannot select more. In the IPACO algorithm, it is better to select the optimal node by the steep ground and the distance, and the meshing is divided according to the minimum sensing radius of the node. When the same barrier gap is repaired, as the number of grids increases, the distance between the repaired barrier gaps gradually decreases (see Fig. 7). This is because the smaller the meshing, the more precise the moving distance, and therefore the smaller the moving distance.

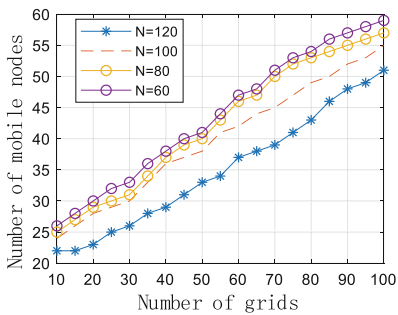


Fig. 6. Number of nodes required

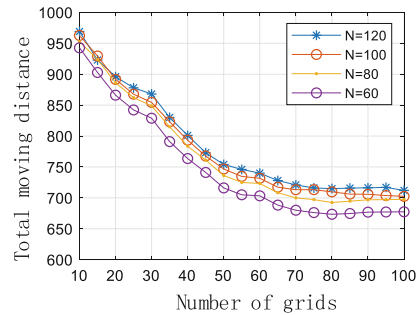


Fig. 7. Average moving distance

5.3 Comparison of Gap Repair Efficiency of Three Algorithms

In the process of barrier repair, the total moving distance of nodes moving to the gap of the barrier is an important indicator to evaluate the advantages and disadvantages of the gap repair method. The shorter the total moving distance, the less energy consumption and the less cost of repairing the barrier. Now assume that the node moves 1 m per unit distance and consumes 3.6 J of energy; The node perception probability is $P = 7.5$. It can be seen that as the number of nodes increases, the total distance of the mobile node to fill the gap decreases, and the total energy consumption also decreases (see Figs. 8, 9, 10 and 11). When the Greedy algorithm is used to repair the barrier gap, the average moving distance of the movable node is the shortest. The total moving distance of the movable node is the shortest when the DBCR algorithm repairs the barrier gap. The total moving distance of the ACO algorithm when repairing the barrier gap is slightly higher than ACO. Because the mobile energy consumption is proportional to the moving distance, it can be seen that the DBCR algorithm is the best, the ACO algorithm is the second, and the Greedy algorithm is the worst. The number of nodes to be repaired and the number of nodes to be moved (see Figs. 9 and 10). The DBCR algorithm is optimal, the ACO algorithm is second, and the Greedy algorithm. This is because the DBCR algorithm introduces a spatial weighting factor when searching for paths, which limits the ants to find paths, so the performance is better. Therefore,

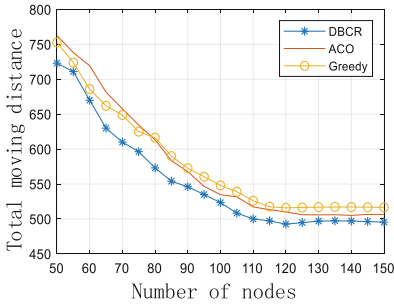


Fig. 8. Comparison of total moving distance

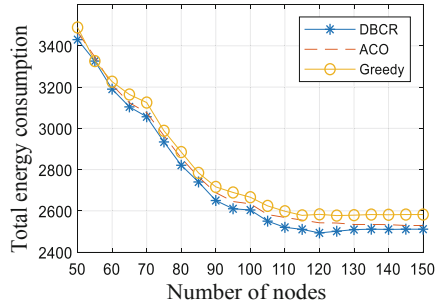


Fig. 9. Comparison of total energy consumption

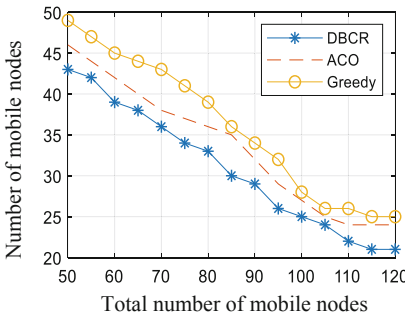


Fig. 10. Comparison of the number of mobile nodes

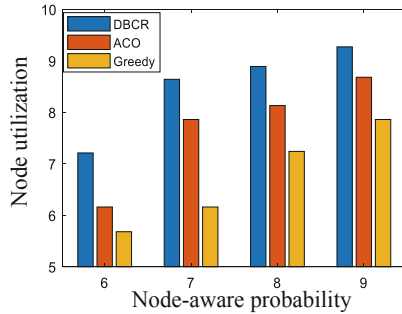


Fig. 11. Comparison of repair rates

experiments have verified that the overall performance of the DBCR algorithm in barrier repair is better than the other two algorithms.

6 Conclusions

Aiming at the problem of barrier gap repair in 3D environment, this paper proposes a distributed barrier gap repair algorithm suitable for 3D environment. Through the improved ant colony algorithm, the repair chain is searched, the relationship model between the mobile node and the repair chain is constructed, and the optimal barrier gap repair position is selected. Finally, the priority distance model of the gap is filled by the moving distance of the node and whether the gap is filled in the key area. Corresponding patching. When using the DBCR algorithm to repair the 3D barrier gap, the number of nodes used can be reduced, and the node moving distance and mobile energy consumption can be reduced.

References

1. Kumar, S., Lai, T.H., Arora, A.: Barrier coverage with wireless sensors. *Wirel. Netw.* **13**(6), 817–834 (2007)
2. Guo, J., Jafarkhani, H.: Movement-efficient sensor deployment in wireless sensor networks (2017)
3. Liu, T., Lin, H., Chen, W., et al.: Chain-based barrier coverage in WSNs: toward identifying and repairing weak zones. *Wirel. Netw.* **22**(2), 523–536 (2016)
4. Si, P., Wu, C., Zhang, Y., et al.: Barrier coverage for 3D camera sensor networks. *Sensors* **17**(8), 1771 (2017)
5. Liu, X.L., Yang, B., Chen, G.L.: Barrier coverage in mobile camera sensor networks with grid-based deployment. *Comput. Sci.* (2015)
6. He, S.: Curve-based deployment for barrier coverage in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **13**(2), 724–735 (2014)
7. Ma, H., Meng, Y., Li, D., et al.: Minimum camera barrier coverage in wireless camera sensor networks. In: *IEEE INFOCOM*, pp. 217–225 (2012)
8. Cheng, C.F., Wang, C.W.: The Target-barrier coverage problem in wireless sensor networks. *IEEE Trans. Mob. Comput.* **17**(5), 1216–1232 (2017)
9. Tao, D., Tang, S., Zhang, H., et al.: Strong barrier coverage in directional sensor networks. *Comput. Commun.* **35**(8), 895–905 (2012)
10. Guvensan, M.A., Yavuz, A.G.: Hybrid movement strategy in self-orienting directional sensor networks. *Ad Hoc Netw.* **11**(3), 1075–1090 (2013)
11. XingGang, F., Chao, W., et al.: A directional K-barrier construction algorithm based on selective box. *J. Comput.* **39**(5) (2016)
12. Dash, D., Gupta, A., Bishnu, A., et al.: Line coverage measures in wireless sensor networks. *J. Parallel Distrib. Comput.* **74**(7), 2596–2614 (2014)
13. Chen, J., Bang, W., Liu, W., et al.: Rotating directional sensors to mend barrier gaps in a line-based deployed directional sensor network. *IEEE Syst. J.* **11**(2), 1027–1038 (2017)
14. Deng, X., Bang, W., Wang, C., et al.: Mending barrier gaps via mobile sensor nodes with adjustable sensing ranges. In: *Wireless Communications and Networking Conference* (2013)
15. Larbi-Mezeghrane, W., Bouallouche-Medjkoune, L., Larbi, A.: Minimum perimeter coverage set based on points of tangency and strong barrier for an extended WSN lifetime. *Wirel. Pers. Commun.* **97**(2), 2339–2358 (2017)
16. Nikitha, K., Rajyalakshmi, D., Damodaram, A.: Effective coverage gap repairing in wireless sensor network. *Int. J. Wirel. Mob. Comput.* **7**(5), 475–484 (2014)
17. Zhao, L., Bai, G., Shen, H., et al.: Energy efficient barrier coverage in hybrid directional sensor networks. In: *International Conference on Wireless Communications and Signal Processing*. IEEE (2015)
18. Wang, Z., Cao, Q., Qi, H., et al.: Cost-effective barrier coverage formation in heterogeneous wireless sensor networks. *Ad Hoc Netw.* **64**, 65–79 (2017)
19. Xiaomin, Z., Ding, F., Keji, M.: An optimization method for WSN barrier gap repair. *Chin. J. Sens. Actuators* (2018)
20. Saipulla, A., Westphal, C., Liu, B., et al.: Barrier coverage with line-based deployed mobile sensors. *Ad Hoc Netw.* **11**(4), 1381–1391 (2013)



Sensor-Cloud Based Precision Sprinkler Irrigation Management System

Mingzheng Zhang, Shuming Xiong, and Liangmin Wang(✉)

Jiangsu University, Zhenjiang 212013, Jiangsu, China
wanglm@ujs.edu.cn

Abstract. The sensor-cloud technology alleviates the restrictions of the traditional wireless sensor networks (WSNs) in terms of storage, computation, and scalability by integrating WSNs with cloud computing. In recent years, sensor-cloud technology is increasingly applied to various real-world applications, especially in agriculture irrigation. With the powerful computing and storage sources, the sensor-cloud enables the massive on-field sensing data to be processed efficiently. Furthermore, the virtualization technology allows multiple clients, typically farmers, to share the same infrastructure resources at a low cost. In this paper, we propose a novel agriculture irrigation system by applying the sensor-cloud technology into the traditional sprinkler irrigation. Targeting the practical irrigation scenes, we illustrate the specific work pattern of the proposed system. Finally, compared with the conventional WSN-based scheme, the simulation results show that our system achieves about 31.06%–41.24% decrease in energy consumption.

Keywords: Sensor-cloud · WSNs · Sprinkler irrigation · Virtualization · Energy consumption

1 Introduction

Sprinkler irrigation has been commonly used for agriculture production in China for a long time due to its low-cost, adaptability, and labor-saving. Various sprinkler irrigation technologies occupy approximately 50% in the market [1]. In the past decade, with the development of the Internet of Things (IoT) technology, WSNs have been widely applied in the agriculture field [2]. Especially in irrigation applications, these technologies have brought new development opportunities to traditional sprinkler irrigation, such as remote monitoring [3, 4], intelligent management [5, 6], and automated irrigation [7, 8]. These advanced technologies have improved the irrigation quality, saved more labor cost, and greatly promoted agriculture production.

However, over the years, due to the excessive use of various sensors, some bottlenecks are gradually emerging, which mainly reflects in two aspects. On the one hand, the network resources of a sensor node such as energy, storage, and computation are limited. So users deploy abundant sensors in monitoring regions,

which results in a massive redundant sensing data, it's difficult for sensor nodes to store and process these data efficiently. Furthermore, duplicate communication among sensor nodes will increase more extra energy consumption. On the other hand, data sharing with others is not easy. Because the sensor nodes are usually deployed by the users only at their interest regions, the third party who has the same demand must deploy their own nodes, leading to a great waste of physical resources. By the way, in the agriculture field, the end users typically are low-income farmers, have difficulty in affording the expenses for deploying and maintaining the system [9].

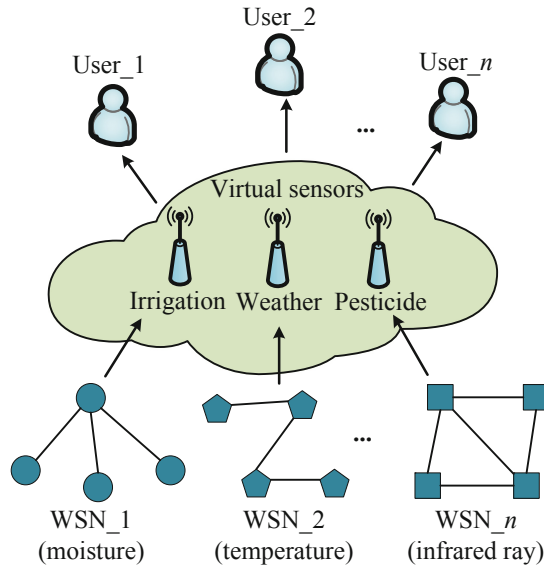


Fig. 1. The basic model of sensor-cloud framework.

Hence, the appearance of sensor-cloud technology has been conceived as a potential solution for the bottlenecks in traditional WSNs. According to the definition in [10], the sensor-cloud is a remote management platform that integrates WSNs with cloud computing. It alleviates the restrictions of traditional WSNs in terms of storage, computation, and scalability by decoupling data producers (i.e., physical sensors) from data providers [11]. In sensor-cloud environment, the complex and energy-consuming tasks are implemented in the cloud, and the physical sensors are mainly responsible for collecting environmental data and send the packets to sink nodes. Moreover, virtualization technology empowers multiple clients to share multiple applications. Figure 1 shows the basic model of the sensor-cloud framework, for instance, the WSN_2 is responsible for collecting the real-time climate data (e.g., temperature) of the farmland, its corresponding weather service is hosted by a virtual sensor created by the sensor-cloud platform. So, users can request the weather service through the relevant virtual sensor.

The sensor-cloud technology conceives a concept of Sensor-as-a-Service (SaaS) [12]. Various physical sensors are provided by multiple sensors owners (SOs) for earning profits. So, farmers have no burden of deploying and maintaining the WSNs. They can request the needed services through web browsers or mobile phones directly without worrying about any other practical problems. Of course, they should pay for usage. By the way, due to the SOs can provide services to multiple users, their profits can also be guaranteed. This multi-tenant, pay-per-use model is conducive to the overall participants in the sensor-cloud framework.

From the above instruction, there are numerous benefits to applying sensor-cloud technology in agriculture field. However, the existing works seldom refer to the management of traditional sprinkler irrigation. So, in this paper, we are committed to reform the traditional sprinkler irrigation by applying sensor-cloud technology. The major contributions of this paper are list in the following.

- (1) We present a novel irrigation system by applying the sensor-cloud technology into the traditional sprinkler irrigation. It can be divided into two subsystems, a sensor-cloud subsystem, and a sprinkler irrigation subsystem. Thus, our system is advanced and low cost.
- (2) We introduce the specific work pattern of each component associated with the proposed system by considering the practical irrigation scenes, which provides a meaningful reference for the design of intelligent sprinkler irrigation system in the future.
- (3) We verify the performance of our scheme through a series of experiments. The results show that the sensor-cloud based scheme is more energy-efficient and inexpensive than the benchmark systems.

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 illustrates the proposed system framework and the specific work pattern in detail. Section 4 verifies the performance of the proposed system. Conclusions and the future work are presented in finally.

2 Related Works

WSNs have been widely applied in various agriculture applications. Irrigation is one of the most important applications, which is known as “Precision Irrigation”. From the definition in [13], the core term of precision irrigation named site-specific management, which means to provide the right amount of water at the right time at the right place. In view of this standard, the authors in [14] design an irrigation management application to monitor the soil moisture using the moisture sensors to optimize the water consumption. In [15], an adaptive irrigation time decision supporting system is designed to reduce water consumption. In another work, the authors in [6] exploit the thermal imaging technology to monitor the temperature change of each region. Thus, the regions that need water can be distinguished if the temperature values are above the threshold value. To achieve remote control of the sprinklers to provide water at the right

places, a kind of valve actuator is designed in [16], which can control up to four valves.

These WSN-based methods are mainly focused on achieving the intelligent and automation of agriculture irrigation. However, the limitations of the sensor nodes in terms of energy, storage, and computation remains an obstacle, which results in several problems, such as the high energy consumption, the massive redundant data, and poor scalability. Hence, over the years, some researchers envision utilizing cloud computing to manage these ubiquitous sensors [17, 18]. The authors in [17] describe the design and the system architecture of sensor-cloud in detail, which provides a platform to manage the physical sensors efficiently. Leveraging the benefits of physical sensors virtualization, numerous users can share the same services via the virtual sensors. In our point of view, the virtual sensors can be regarded as logical sensors generated by multiple physical sensors [19]. The authors in [20] have introduced four virtual sensors configurations in detail, such as one-to-many, many-to-one, many-to-many, and derived.

As for applications in the agriculture field, the authors in [21] present an agriculture sensor-cloud infrastructure to provide various agricultural services. But they pay all attention to the routing protocol in the physical layer. Thereafter the authors in [9] analyze the benefits of applying the sensor-cloud framework for agricultural. However, they haven't introduced the specific work process. Then in [22], a sensor-cloud based M2M (measurement to management) system for precision irrigation is proposed, and in order to reduce the energy consumption, the mobile sensor robots are exploited to collect environment data. However, this method is difficult for general farmers to use in rural areas.

3 Proposed System Architecture

In this section, we introduce the proposed irrigation sensor-cloud system architecture in detail. The whole system can be divided into two subsystems: a sensor-cloud subsystem, and a sprinkler irrigation subsystem.

3.1 Sensor-Cloud Subsystem

As shown in Fig. 2, the sensor-cloud subsystem can also be divided into three layers: the user layer, the middleware layer, and the physical layer.

In the user layer, multiple users can share the same sensor-cloud infrastructure. The sensor-cloud platform provides users several interfaces so that users can query the on-field information and request services via web browsers or mobile phones. Notice that different users usually have different authorizations. For instance, some business organizations or government institutions can request on-field soil moisture data. However, only the farmers can start the irrigation program.

The task of the middleware layer is to configure virtual sensors. The virtual sensors play the role of data providers in sensor-cloud. For each irrigation region, its environment data is stored in the corresponding virtual sensors, which are

created by multiple physical sensors in this region. From our point of view, the virtual sensors can be regarded as the logical mapping of a set or subset physical sensors. When farmers want to know whether their farmland needs irrigation, they can request the corresponding virtual sensors service, if the current soil moisture below the threshold value and there is no rain in the next 12 h. Then they can start the irrigation program. In this process, the processing, analysis, and visualization of sensing data are all done by the middleware.

The physical layer includes various sensors (e.g., moisture, temperature, and humidity) and numerous irrigation devices (e.g., valve actuator, sprinkler, and pump). Both of them are deployed by device owners (DOs) for earning profits. The sensors are responsible for collecting the real-time environment data and send to the sink node, and the irrigation devices are responsible for implementing irrigation decisions. All of the sensors and devices are efficiently managed in the sensor-cloud platform, even if they are in different regions or belong to different farmers.

3.2 Sprinkler Irrigation Subsystem

As shown in Fig.3(a), the sprinkler irrigation subsystem is also part of the physical layer.

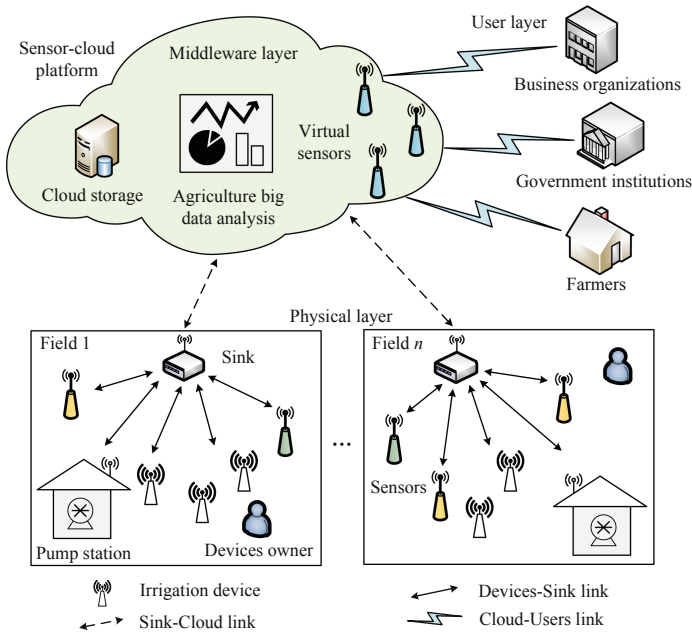


Fig. 2. The proposed irrigation sensor-cloud framework.

The pump station is applied for pumping water and providing sprinklers a suitable operating pressure. It works according to the instruction of the controller. The valve actuator can control the states (i.e., on or off) of sprinklers via the solenoid valves, which are the most frequently used device in irrigation. The function of sprinkler is to spray water to the field and ensure the irrigation uniformity. It usually has an optimal operating pressure range, which can be measured by pressure gage [23].

Then, as shown in Fig. 3(b), we choose rectangle mode to deploy sprinklers. The whole field is divided into multiple sub-areas (e.g., $ABCD$) by every two neighboring sprinklers. The sensor-cloud platform will create a virtual sensor for each sub-area to store the corresponding sensing data. We use the matrix to visually indicates the state of sub-areas and sprinklers.

$$A_{mn} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \quad (1)$$

For $\forall 1 \leq i \leq m, 1 \leq j \leq n, a_{ij} = 0$ indicates the soil moisture level in a_{ij} is above the threshold, otherwise $a_{ij} = 1$. Furthermore, for $\forall a_{ij}$, there are four sprinklers and can be indicated by matrix B_{ij} .

$$B_{ij} = \begin{bmatrix} b_{ij}^1 & b_{ij}^2 \\ b_{ij}^3 & b_{ij}^4 \end{bmatrix} \quad (2)$$

Similarly, $b_{ij}^k = 0$ ($k = 1, 2, 3, 4$) indicates the sprinkler is closed, $b_{ij}^k = 1$ indicates the sprinkler is open.

4 Work Pattern

Some previous studies have defined the basic virtualization model of sensor-cloud. However, they haven't involved specific applications. In this section, we refine the definitions of the relevant components and introduce the specific work pattern by taking into account the practical sprinkler irrigation scenes. In our scheme, we divide all the components into three parts: service provider, decision maker, and executor.

4.1 Service Provider

The service provider mainly includes platform owners and device owners.

- (1) *Platform owner*: The platform owner is the administrator that manages the sensor-cloud services and provides farmers various interfaces that allow them to access soil moisture data at any time. Besides the irrigation service, there are also some other applications, such as weather service, pesticide service, etc. Moreover, the administrator can formulate the charge standard

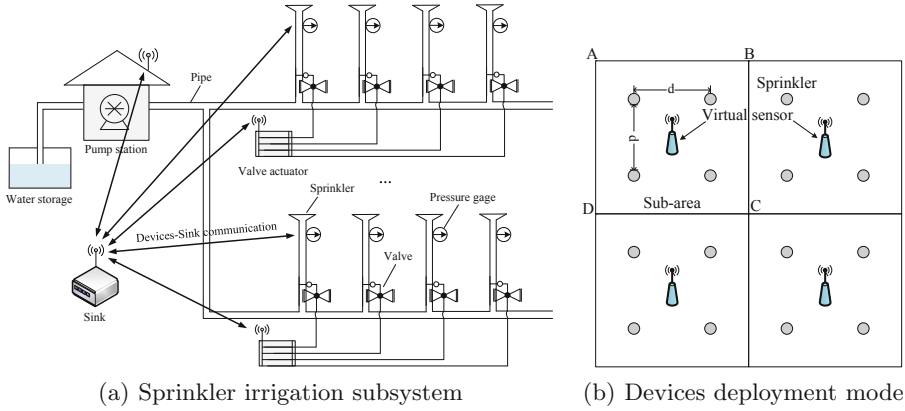


Fig. 3. The sprinkler irrigation system.

and charge for these services. Here, a platform owner (PO_i) can be defined as:

$$PO_i = \{PO_s, PO_c, PO_i, PO_{area}, PO_\Psi\} \tag{3}$$

where PO_s is the provided services, PO_c is the used cloud service, PO_i is the provided interfaces, PO_{area} is the service area, and PO_Ψ is the charge standard.

- (2) *Device owner*: The device owners are providers of physical sensors and other irrigation devices. A device owner can be a business organization, a government institution, or a farmer. Of course, different roles have different authorizations. These sensors or devices can be invoked through the virtual sensors. For each device owner ($DO_i \in DO$), there is:

$$DO_i = \{DO_{id}, DO_r, DO_s, DO_{area}\} \tag{4}$$

where DO_{id} and DO_r are the owner’s identifier and role, DO_s is the relevant service, and D_{area} is the deployment area.

- (3) *Service model*: Similarly, the service model (S_i) can be expressed as:

$$S_i = \{S_{type}, S_{area}, S_t, S_f, S_d, S_\Psi\} \tag{5}$$

where S_{type} , S_{area} , and S_t are the service type, area, and time, respectively. S_f is the farmer who requests the service, S_d is the requested devices, and S_Ψ is the generated expense.

4.2 Decision Maker

The decision maker mainly includes the end users and the middleware in the sensor-cloud platform.

- (1) *End user*: The end users, usually farmers, can request virtual sensor services to satisfy their demand. They can control the data collection interval of virtual sensors via the web browser or mobile phone, decide the irrigation time, area, and used sprinklers. When the virtual sensors are not necessary, the farmers can log out at any time. They don't need to know more details about the physical sensors. Thus, for an end user $U_i \in U$, the definition can be denoted as:

$$U_i = \{U_{id}, U_{role}, U_{area}, U_{\psi}\} \quad (6)$$

where U_{id} and U_{type} are the user's identifier and role, respectively. U_{area} is the farmland that belongs to U_i (if U_i is a farmer), and U_{ψ} is the user's authorization.

- (2) *Middleware*: When the user's request arrives, the middleware should decide which physical sensors to choose to configure the virtual sensors. Traditional WSN tends to exploit all the deployed physical sensors to collect data. However, in sensor-cloud, only a subset of physical sensors is selected to create virtual sensors in response to user requests. Here, a virtual sensor (V_i) can be defined as:

$$V_i = \{V_{id}, V_{area}, V_s, V_u, V_p\} \quad (7)$$

where V_{id} , V_{area} , and V_s indicate the virtual sensor's identifier, relevant farmland, and hosted service, respectively. V_u is the user who requests service, V_p is the used physical sensors.

- (3) *Decision model*: Similarly, a decision or a service request (R_i) can be defined as:

$$R_i = \{R_{role}, R_{area}, R_s, R_d, R_t\} \quad (8)$$

where R_{role} is the role of the decision maker, R_{area} is the interested area, R_s and R_d are the requested service and devices, and R_t is the service time.

4.3 Executor

The executor usually includes the on-field physical sensors and various irrigation devices.

- (1) *Physical sensor*: The physical sensors (P) are responsible for collecting environmental data and creating virtual sensors to provide the corresponding services. For a physical sensor $P_i \in P$, the definition can be denoted as:

$$P_i = \{P_{id}, P_{type}, P_{state}, P_{area}, P_d\} \quad (9)$$

where P_{id} , P_{type} , P_{state} , and P_{area} are the physical sensor's identifier, type, state (e.g., active or dormant), and deployed area, respectively. P_d is the device owner.

- (2) *Irrigation device*: The irrigation devices are responsible for implementing the farmer's irrigation decisions. These devices work according to the instructions of the micro controller. So, for an irrigation device $D_i \in D$, the definition can be expressed as:

$$D_i = \{D_{id}, D_{type}, D_{state}, D_{area}, D_d, D_{\psi}\} \quad (10)$$

where D_{id} , D_{type} , D_{state} , D_{area} , and D_d denote the device's identifier, type, state, installed area, and owner, respectively. D_{ψ} is the property of the device (e.g., pump power, sprinkler operating pressure).

5 Performance Evaluation

5.1 Simulation Settings

To evaluate the performance of the proposed system, we perform simulations using NS-3.28 in Ubuntu16.04. Table 1 enlists the used parameters. In the simulation, we compare our scheme with the conventional WSN in terms of energy consumption, network lifetime, and cost. For both schemes, the physical sensors are homogeneous and distributed randomly in the simulation area (100×100), and the energy consumption is associated with distance and package size. The difference is the communication method. In the traditional WSN scenario, the LEACH protocol is applied. In the sensor-cloud scenario, the nodes communicate directly with the sink node (located at 50, 100). The initial energy of each node is $0.1 j$ (set the energy low, so it's easier to plot and manage the raw data).

Table 1. Simulation parameters.

Parameter	Value
Simulation area (m^2)	100×100
Number of nodes	50–100
Message size (<i>bits</i>)	256
Nodes transmission range (m)	100
Initial node energy (j)	0.1

5.2 Evaluation Metrics

We analyze the performance of two schemes with respect to the following metrics:

- (1) *Energy consumption*: The energy consumption (E) denotes the total energy consumed per transmission period, the value of E can be calculated by:

$$E = \sum_{p=1}^n (E_p^t + E_p^r + E_p^s + E_p^c) \quad (11)$$

where n is the number of nodes, E_p^t , E_p^r , E_p^s , and E_p^c are the energy consumption due to transmission, receiving, sensing, and computing, respectively.

- (2) *Network lifetime*: For both of the WSN and sensor-cloud, the network lifetime (T) indicates the time that the service can be provided. We use the rounds of iterations when the overall residual energy arrives the threshold to evaluate this metric, which can be expressed as:

$$T = R_t \tag{12}$$

where R_t denotes the rounds of iterations when the overall residual energy reaches the threshold.

- (3) *Cost*: From the farmer’s point of view, the cost in sensor-cloud is the rental of the physical sensors and irrigation services, which can be defined as:

$$C_{sc} = n\gamma \times r_{sc} + r_{is} \tag{13}$$

where n is the total number of physical sensors, γ is the utilization rate of the physical sensors, r_{sc} is the unit price of each sensor, and r_{is} is the unit price of irrigation service.

As for in traditional WSN, the cost can be defined as:

$$C_{wsn} = n \times c_{dep} + n\beta \times c_{dep} + c_{is} + r_{is} \tag{14}$$

where n' is the total number of sensor nodes, c_{dep} is the cost of deploying the system, β is the fault rate, c_{is} is the cost of irrigation system.

5.3 Results and Analysis

First, we adopt the value of energy consumption when the number of iterations reaches 1000. To reduce the deviation, we calculate the average of 10 simulations. The results are shown in Fig. 4.

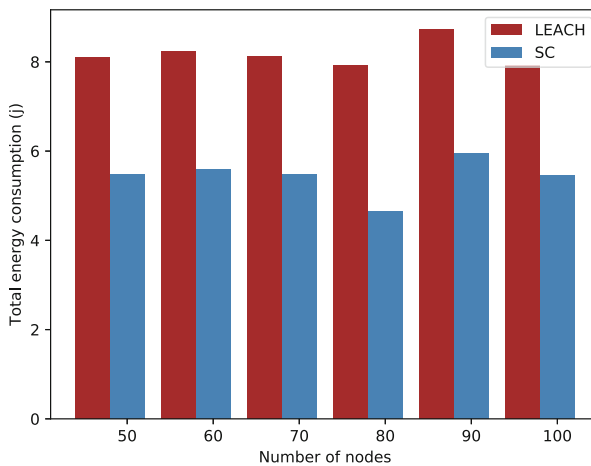


Fig. 4. Energy consumption with different network size.

In terms of transmission, receiving, sensing, and computing, there is a large amount of redundant communication between sensor nodes in traditional WSN, which consumes numerous extra energy. However, in the sensor-cloud environment, only a part of physical sensors are utilized, and the sensor-cloud platform has the globe view of the network, the data can be forwarded to sink node through the shortest path. So there is nearly no redundant communication. The results show that the sensor-cloud achieves about 31.06%–41.24% decrease in energy consumption.

Then we adopt two different network sizes, 50 nodes, and 100 nodes, to evaluate the network lifetime. The threshold of residual energy is set at 10%. In the traditional WSN, there are about 600 rounds. However, in sensor-cloud, it can reach 2000 rounds at 50 nodes, and more than 2500 rounds at 100 nodes. The results are shown in Fig. 5.

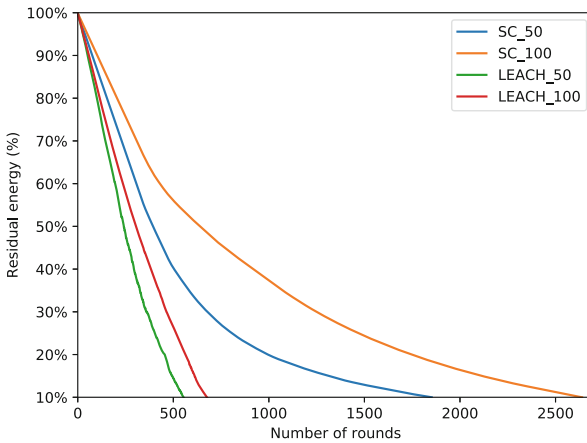


Fig. 5. Network lifetime of two scenarios.

Finally is the cost, we assume the following unit price values: $r_{sc} = 1$, $r_{is} = 1$, $c_{dep} = 3$, $p = 0.3$, and $c_{is} = 100$. Thus, as shown in Fig. 6(a), the cost in traditional WSN is proportional to the network size. However, in the sensor-cloud framework, the cost is almost the same regardless of the size of the network. That's due to not all physical sensors are utilized to reply to the user's request. In fact, for a farmer, the cost is only related to the rental time. So, we assume that the network size is fixed and the system will be maintained once a month. The relationship between cost and time is shown in Fig. 6(b). We assume that farmers have been renting services without stopping. Even so, it will take 15 months to make the cost the same.

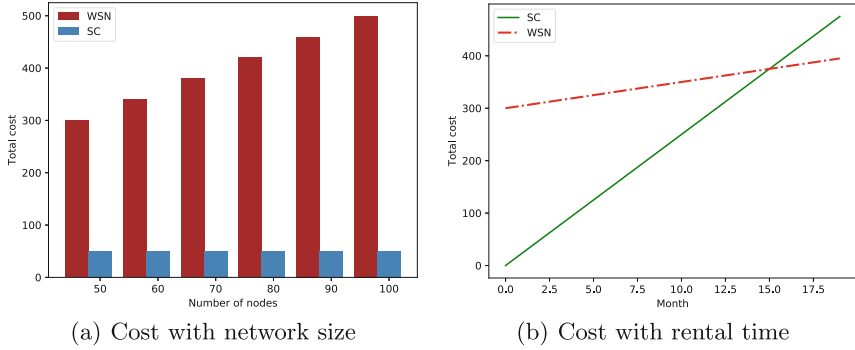


Fig. 6. Comparison of cost in two scheme.

6 Conclusions

The existing intelligent agriculture applications are complex and high cost. For a general farmer, usually has difficult to afford the cost of deploying and maintaining the system. The appearance of sensor-cloud technology provides an alternative solution for the future development of agriculture modernization. The farmers don't need to know more details about the physical sensors or other practical problems. Furthermore, leveraging the benefits of virtualization technology, multiple farmers can share the same sensor-cloud infrastructure with low cost.

In this paper, we are committed to the reform of the traditional sprinkler irrigation system and extend it to the sensor-cloud framework. We introduce the overall system architecture and the specific work pattern of each competent in sensor-cloud. Furthermore, we analyze the performance of the proposed system compared to the traditional WSN concerning energy consumption, network lifetime, and cost. The experiment results show that our scheme reduces about 31.06%–41.24% energy consumption, and saves a massive amount of cost for farmers.

Nonetheless, there are still some issues that need further elaboration in future studies, such as virtual sensor provisioning, more detailed price model for all participants, and data security.

References

1. Li, Y., Bai, G., et al.: Development and validation of a modified model to simulate the sprinkler water distribution. *Comput. Electron. Agric.* **111**, 38–47 (2015)
2. Elijah, O., Rahman, T.A., et al.: An overview of Internet of Things (IoT) and data analytics in agriculture: benefits and challenges. *IEEE Internet Things J.* **5**(5), 3758–3773 (2018)
3. Corke, P., Wark, T., et al.: Environmental wireless sensor networks. *Proc. IEEE* **98**(11), 1903–1917 (2010)

4. Malaver, A., Motta, N., et al.: Development and integration of a solar powered unmanned aerial vehicle and a wireless sensor network to monitor greenhouse gases. *Sensors* **15**(2), 4072–4096 (2015)
5. Goap, A., Sharma, D., et al.: An IoT based smart irrigation management system using machine learning and open source technologies. *Comput. Electron. Agric.* **155**, 41–49 (2018)
6. Roopaei, M., Rad, P., et al.: Cloud of things in smart agriculture: intelligent irrigation monitoring by thermal imaging. *IEEE Cloud Comput.* **4**(1), 10–15 (2017)
7. Nikolidakis, S., Kandris, D., et al.: Energy efficient automated control of irrigation in agriculture by using wireless sensor networks. *Comput. Electron. Agric.* **113**, 154–163 (2015)
8. Sudha, M.N., Valarmathi, M., et al.: Energy efficient data transmission in automatic irrigation system using wireless sensor networks. *Comput. Electron. Agric.* **78**(2), 215–221 (2011)
9. Ojha, T., Misra, S., et al.: Sensing-cloud: leveraging the benefits for agricultural applications. *Comput. Electron. Agric.* **135**, 96–107 (2017)
10. Alamri, A., Ansari, W.S., et al.: A survey on sensor-cloud: architecture, applications, and approaches. *Int. J. Distrib. Sens. Netw.* **9**(2), 917–923 (2013)
11. Dinh, N., Kim, Y.: An energy efficient integration model for sensor cloud systems. *IEEE Access* **7**, 3018–3030 (2018)
12. Misra, S., Chatterjee, S., et al.: On theoretical modeling of sensor cloud: a paradigm shift from wireless sensor network. *IEEE Syst. J.* **11**(2), 1084–1093 (2014)
13. Chen, N., Zhang, X., et al.: Integrated open geospatial web service enabled cyber-physical information infrastructure for precision agriculture monitoring. *Comput. Electron. Agric.* **111**, 78–91 (2015)
14. Navarro, H.H., Torres, S.R., et al.: A wireless sensors architecture for efficient irrigation water management. *Agric. Water Manag.* **15**, 64–74 (2015)
15. Fazackerley, S., Lawrence, R.: Reducing turfgrass water consumption using sensor nodes and an adaptive irrigation controller. In: 2010 IEEE Sensors Applications Symposium (SAS), pp. 90–94. IEEE, Limerick (2010)
16. Coates, R.W., Delwiche, M.J., et al.: Wireless sensor network with irrigation valve control. *Comput. Electron. Agric.* **96**, 13–22 (2013)
17. Yuriyama, M., Kushida, T.: Sensor-cloud infrastructure-physical sensor management with virtualized sensors on cloud computing. In: NBiS, vol. 10, pp. 1–8 (2010)
18. Dwivedi, R.K., Kumar, R.: Sensor cloud: integrating wireless sensor networks with cloud computing. In: 2018 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON), pp. 1–6. IEEE, Gorakhpur, India (2018)
19. Lim, Y., Park, J.: Sensor resource sharing approaches in sensor-cloud infrastructure. *Int. J. Distrib. Sens. Netw.* **10**(4), 1–8 (2014)
20. Madria, S., Kumar, V., et al.: Sensor cloud: a cloud of virtual sensors. *IEEE Softw.* **31**(2), 70–77 (2013)
21. Kim, K., Lee, S., et al.: Agriculture sensor-cloud infrastructure and routing protocol in the physical sensor network layer. *Int. J. Distrib. Sens. Netw.* **10**(3), 1–13 (2014)
22. Tyagi, S., Obaidat, M.S., et al.: Sensor cloud based measurement to management system for precise irrigation. In: GLOBECOM 2017–2017 IEEE Global Communications Conference, pp. 1–6. IEEE, Singapore (2018)

23. Salvatierra, B.B., Montero, M., et al.: Development of an automatic test bench to assess sprinkler irrigation uniformity in different wind conditions. *Comput. Electron. Agric.* **151**, 31–40 (2018)
24. Vuran, M.C., Akan, O.B., et al.: Spatio-temporal correlation: theory and applications for wireless sensor networks. *Comput. Netw.* **45**(3), 245–259 (2004)
25. Lemos, M., Rabelo, R., et al.: An approach for provisioning virtual sensors in sensor clouds. *Int. J. Netw. Manag.* **29**(2), 1–21 (2019)



Deep Memory Network with Auxiliary Sequences for Chinese Implied Sentiment Analysis

Chao Wang¹, Yunhua He^{1(✉)}, Limin Sun², Chengjie Pang¹, and Jitong Li¹

¹ North China University of Technology, Beijing 100144, China
{wangchao.andy,yunhuahe,chengjiepang,jitonglig}@ncut.edu.cn

² University of Chinese Academy of Sciences, Beijing 100093, China
sunliming@iie.ac.cn

Abstract. Sentiment analysis is a hot topic and has various application scenarios. The polarity recognition of implied sentiment in a sentence can be achieved by the way of statistic and prediction. However, the polarity of sentiment is influenced by funny, humorous, and ironic Internet cultures, therefore it is hard to be verified. In this paper, we use a deep memory network with the auxiliary sequence to obtain the text feature vectors. Then the Emoji set and the special word set from the internet are imported, which are combined with the formal text feature vectors to form the classification feature vectors. At last a binary classifier is designed to get the final polarity prediction. Besides, an incremental online learning method with feedback adjustment is introduced to update the Emoji set and the special word set. Experiment results show that, on the IMDB datasets the prediction accuracy is about 85% and on the Chinese implied sentiment evaluation datasets the prediction accuracy is about 96%, which prove the effectiveness of the model.

Keywords: Deep memory network · Auxiliary sequence · Implied sentiment analysis

1 Introduction

Currently, sentiment analysis is a hot topic in both industry and academia. Given some negative, neutral or positive messages, sentiment recognition is to identify the emotional polarity of the target text. According to the text granularity, the sentiment classification can be divided into four levels – word-level, aspect-level, sentence-level, and chapter-level. The main task of establishing a classification model is to obtain available feature vector expressions of target contexts. Most of proposed solutions are based on supervised machine learning approaches [10].

Supported by the National Natural Science Foundation of China (No. 61802004, 61802005), Natural Science Foundation of Beijing (No. 4184085), Startup Foundation of North China University of Technology, Scientific Foundation for Yuyou Talents.

However, deep learning has produced extremely promising results for various tasks on natural language understanding, particularly on topic classification, sentiment analysis, question answering and language translation area [2, 12, 18].

However, influenced by funny, humorous, and ironic Internet cultures, the polarity of sentiment is not always shown literally. In the Internet culture, there is a kind of expression called Emoji. When the same sentence has different Emoji expressions, the sentiments they express are not same, even opposite. Bjarke [6] uses millions of Emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. For example, In sentence “The keyboard man in the review is really superb [ridiculous]”, the sentiment polarity seems to be positive on the literal meaning, however in fact it should be negative due to the ironical use of the [ridiculous] expression. The same problem can also be found in some special words. For example, in the Chinese sentence “The green hat on his head is really good-looking”, the special word “green hat” conceals extremely negative sentiment in Chinese culture. Therefore, the research on implicit sentiment recognition is another research topic, which has not attracted much attention. Most of researchers only focus on the ironic recognition scene. Aditya [9] presents a computational system that harnesses context incongruity as a basis for sarcasm detection. As far as it is concerned, it is of great research value, because they usually carry very subtle, effective, and strongly infective feedback information.

In this paper, based on auxiliary sequence with inference function, a deep memory network combined with the Emoji set and a special word set is used to achieve the polarity classifications of implicit sentiment. The proposed model obtains the feature vectors of texts, Emoji expressions and the special words (such as green hat so on) by the way of supervised learning. Then, the binary classifier is designed to predict polarity by extending dimensions to the final classification feature vectors. On the IMDB datasets, the verification accuracy is about 85%, which proves the theoretical feasibility of the model. However, since there is no open source Chinese implied sentiment evaluation datasets with Emoji and special words, we capture data through the Internet and the verification accuracy is about 96%, which proves the effectiveness of the model.

The following are the three statements of this paper: (1) The model has the inference function by introducing the auxiliary sequence, so that the same word vector can be inferred to different mathematical vector representations in different semantic environments. (2) Inference function also makes memory continuously update and optimize in feedback training. (3) The vector representations of the special word set and the Emoji expression are constructed and updated by an incremental online learning method with negative feedback adjustment, so that the prior knowledge with implied sentiment polarity is introduced in order to get better recognition results through the classifier.

2 Related Work

2.1 Attention Mechanisms

In 2014, Graves et al. [8] extend the capabilities of neural networks by coupling them to external memory resources, which they can interact with by attentional processes. They call their device “Neural Turing Machine” (NTM) whose architecture is differentiable end-to-end and which can be trained with gradient descent.

In the Neural Turing Machine architecture, the attention mechanisms are interpreted as the addressing mechanisms which are used to produce the weightings. These weightings rise by combining two addressing mechanisms with complementary facilities. The first mechanism – content-based addressing– focuses on locations based on the similarity between their current values and values emitted by the controller. The second is location-based addressing. The current relative location information is introduced as a calculation factor to generate weights. Usually, the two mechanisms are applied to the vast majority model at the same time.

Based on the above, Tang and Qin [15] introduce a deep memory network for aspect level sentiment classification. They implement different attention strategies and show that leveraging both content and location information could learn better context weight and text representation. They also demonstrate that using multiple computational layers in memory network could obtain better performance.

2.2 Memory Network

In 2014, Weston [16] introduced a new class of learning models called memory networks. Memory networks make inference by an inference component combined with a long-term memory component. The long-term memory can be read and written to predict something. Generally, a memory network consists of an array of objects called memory M and four components called I , G , O and R , where I converts input to internal feature representation, G updates old memories, O generates an output representation and R outputs a response.

Based on their work, Sukhbaatarb et al. [14] propose a neural network with a recurrent attention model over a possibly large external memory. Unlike previous model, their model is trained end-to-end, and hence requires significantly less supervision during training. Their model approaches the same performance with previous model and is significantly better than other baselines with the same level of supervision.

Dou et al. [4] propose a deep memory network for document-level sentiment classification which could capture the user and product information at the same time. They conduct experiments on IMDB and Yelp datasets and the results prove the effectiveness of their algorithms.

2.3 Online Learning

Deep learning cannot be satisfactorily achieved without real-time updating. Online learning must create challenging activities that enable learners to link new information to old [1]. Duchi et al. [5] present a new family of subgradient methods that dynamically incorporate knowledge of the geometry of the data observed in earlier iterations to perform more informative gradient-based learning. And they give several efficient algorithms for empirical risk minimization problems with common and important regularization functions and domain constraints.

3 Proposed Methods

3.1 Model Definition

Based on the attention mechanism in memory networks, an Emoji set and a special words set are imported into a deep memory network with auxiliary sequence. Its architecture consists of six parts, including embedding layer, auxiliary sequence, attention layer, Emoji and special word processing component, feature fusion classifier and online learning module. The six parts will be described in Sect. 3.2. The specific model architecture diagram is shown in Fig. 1.

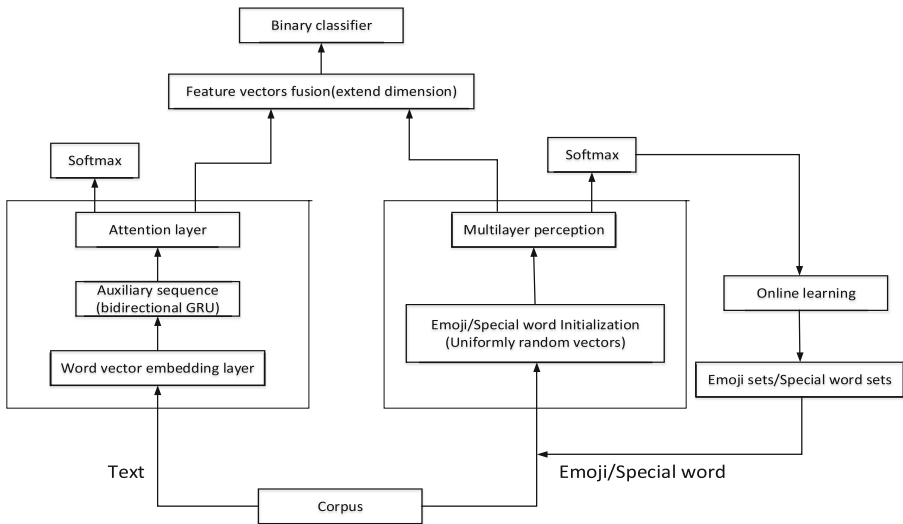


Fig. 1. Architecture of model

The workflow of the model is as follows. When the corpus contains Emoji, the Emoji is extracted first. Then the text is mapped to the vector space which

is the input of the auxiliary sequence (bidirectional GRU), thus obtaining the implicit state of the auxiliary sequence as memory. The attention mechanism is applied to the generated memory, and the text feature vectors is obtained through supervised training. At the same time, according to the Emoji set and the special word set, the extracted Emoji and the retrieved special words are initialized to the unified random vectors as the input of the processing components of the MLP. Then the supervised training is carried out in order to obtain the feature vectors of the Emoji and special word set. At last, the feature fusion component fuses text feature vectors, Emoji feature vectors and special word feature vectors through a method of dimension extending. And the final vectors are input to the classifier, which is used to predict the polarity of implicit sentiment.

In the Emoji set and special words set, there are entities for representing null values. They are used to process sentences which include neither Emoji nor retrieved special words in order to uniformly train and test corpus from end to end. In addition, when the sentence contains more than one Emoji or special word, their equal weight linear average values are calculated separately so that the model sizes can be unified.

3.2 Model Framework

Embedding Layer. Sentences in the corpus are the input of the model. Each of these sentences consists of n words, denoted by $s = \{\omega_1, \omega_2, \dots, \omega_i, \dots, \omega_n\}$ (s is the sentence and ω_n is the n -th word). When the model processes corpus texts, they need to be mapped to a low-dimensional continuum of real-valued vector spaces referred as “word embedding”. By the mapping relationship, the embedded vector of each sentence can be obtained, $e = \{e_1, e_2, \dots, e_i, \dots, e_n\}$. For the entire text corpus, it can be represented by a matrix. We denote text with the matrix L , $L \in \mathbb{R}^{d \times |V|}$, where d is the dimension of the word vector and $|V|$ is the size of the vocabulary. Their relationship is as follows:

$$e_i = W_e \cdot \omega_i, (i \in [1, n]) \quad (1)$$

W_e is the matrix representation of the mapping, and referred as the embedding matrix. i denotes the i -th word.

Auxiliary Sequence. The auxiliary sequence with certain inference function is used as a bridge to connect embedded vectors with memory. In the implicit sentiment polarity classification scene, the same word may have different or opposite effects on polarity on different contexts. Therefore, the memory of the word should be updated in real time according to the specific context to obtain better classification results. The bidirectional GRU [3] sequence model performs iterative training using the supervised signal of implied sentiment polarity, and infers the contribution of this word to the polarity of implicit sentiment from the context of each word’s embedded vector bidirectionally.

The specific calculation flow is as follows:

$$\mathbf{h}_t = GRU(e), t \in [1, T] \quad (2)$$

$$\mathbf{h} = \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_t, \dots, \mathbf{h}_T \quad (3)$$

$$\mathbf{m} = (\mathbf{h}_+, \mathbf{h}_-) \quad (4)$$

In the equation, t denotes the t -th element in the sequence model, and h_t denotes the implicit state of the t -th element in the sequence. h_+ and h_- denote the implicit state of the entire forward sequence and the hidden state of the entire reverse sequence respectively. \mathbf{m} represents memory, which is formed by combining the entire hidden states of the forward and backward sequences. Besides, there are $2T$ memory cells.

Attention Layer. The attention layer is used to generate corresponding weight for each memory cell. According to the model's requirement, it can be stacked in multiple layers, and the high-level features can be gradually expressed by the underlying feature abstraction. Due to the characteristics of parameter sharing and linear characterization in the attention mechanism, there will be less parametric explosion phenomena when multiple layers are stacked. Each memory unit has different contributions to text feature vectors. Using ordinary feed-forward network, each memory cell (represented by $[m_{t'}, t' \in [1, 2T]]$) is transformed into a vector $u_{t'}$ which is used as the attention level to calculate the weight. At the same time, the vector u_ω of word level is introduced to calculate similarity with $u_{t'}$, and the weight factor $\alpha_{t'}$ is generated by the softmax function. Additionally, the feature vector of the text is linearly weighted by the memory unit. u_ω can be seen as a high-level representation [11]. It can be initialized randomly and learned jointly during training [17]. The specific calculation flow is as follows:

$$\mathbf{u}_{t'} = \tanh(\mathbf{W}_\omega \mathbf{m}_{t'} + \mathbf{b}_\omega) \quad (5)$$

$$\alpha_{t'} = \frac{\exp(\mathbf{u}_{t'}^T \mathbf{u}_\omega)}{\sum_{t'=1}^{2T} \exp(\mathbf{u}_{t'}^T \mathbf{u}_\omega)} \quad (6)$$

$$\mathbf{s} = \sum_{t'=1}^{2T} \alpha_{t'} \mathbf{m}_{t'} \quad (7)$$

In the above equations, \mathbf{s} is the output vector of the attention layer, which is the final feature vector of the corpus text (if the attentional layer consists of multiple layers, \mathbf{s} can be used as the input of the following attentional layer). Then the final output vector \mathbf{s} of the attention layer is sent to the softmax layer, so that the supervision training can be carried out. Therefore, the feature vector of the text is obtained through the deep memory network based on the auxiliary sequence and the supervised training.

Emoji and Special Word Processing Component. According to the Internet culture and special cultural background, a special word set is imported for retrieving in corpus texts. Once a word contained in the special word set is found in the text, the word is recorded and initialized with a random vector. At the same time, the Emoji emoticons extracted before are also initialized with random vectors. According to the idea of Word2Vec, with the polarity of implicit emotions as the feedback signal for supervised training, the final special words and vector representations of Emoji expressions are obtained through the multi-layer perceptron (MLP).

The specific calculation flow is as follows.

$$\mathbf{u}_{Emoji} = \arg\{\text{softmax}(\tanh(\mathbf{W}_E \mathbf{u}_{Emoji} + \mathbf{b}_E))\} \quad (8)$$

$$\mathbf{u}_{special} = \arg\{\text{softmax}(\tanh(\mathbf{W}_s \mathbf{u}_{special} + \mathbf{b}_s))\} \quad (9)$$

\mathbf{u}_{Emoji} denotes Emoji feature vector. $\mathbf{u}_{special}$ denotes the special word feature vector. $\arg(*)$ denotes the input vector that makes the MLP prediction accuracy highest. According to the specific situation, the above formulas can be regularized.

Feature Fusion Classifier. Once the text feature vector \mathbf{s} , the Emoji feature vector \mathbf{u}_{Emoji} and the special word feature vector $\mathbf{u}_{special}$ in the corpus are obtained, feature fusion is performed and a binary classifier is designed to recognize the implicit sentiment. The feature fusion method in this paper is the direct dimension expansion. Based on GBDT [7], the classifier is constructed to evaluate the overall model. The specific calculation flow is as follows.

$$\mathbf{X} = (\mathbf{s}, \mathbf{u}_{Emoji}, \mathbf{u}_{special}) \quad (10)$$

$$c = \arg \max P(\mathbf{X}) \quad (11)$$

where \mathbf{X} denotes the input vector of the classifier, c denotes the final prediction category, and P denotes the probability prediction function.

The gradient-descent algorithm is used to optimize the cross-entropy loss values with softmax. The specific calculation process is as follows.

$$\text{loss} = - \sum_{(s) \in S} \sum_{c \in C} P_c^g(s) \cdot \log(P_c(s)) \quad (12)$$

$$c = \arg \max P_c(s) \quad (13)$$

S represents the training set including all sentences, C represents a set of all categories. $P_c(s)$ indicates the prediction probability that the text feature vector s belongs to c class. $P_c^g(s)$ is the tag value of training set with ‘1’ indicating that s belongs to c class and ‘0’ not.

Online Learning Module. Based on the set of words and the TF-IDF [13] algorithm, high-frequency words and Emoji symbols in classified samples can be used to help correct the classification result. The set of Emoji and special words are initialized based on the current language and network culture respectively. For each item in the set, its coefficient on the effect of implied positive and negative sentiment is set. By setting a certain threshold, new words and Emoji that carry implicit sentiment tendencies can be adjusted to influence coefficient and can be excavated from high-frequency words and Emoji in the set of words built with TF-IDF. In this way, through the feedback adjustment mechanism, the model continuously conducts online learning so as to expand and update the special word set and the Emoji set in real time. The following algorithm provides an online learning extension and update function for special word sets based on feedback.

Algorithm 1. Incremental online learning method with feed-back adjustment

Input:

- The set of positive samples for current batch, P_n ;
- The set of negative samples for current batch, N_n ;
- The set of model prediction results for current batch, R_n ;
- The set of Emoji on former batches, E_{n-1} ;
- The set of special words on former batches, S_{n-1} ;
- The value of frequency threshold, a ;
- The learning rate, e ;

Output:

- The set of Emoji for current batch, E_n ;
 - The set of special words for current batch, S_n ;
-

1. According to the TF-IDF algorithm, the bag of words are build iteratively accumulated;
2. Based on the prediction of R_n , traverse P_n and N_n ;
3. Filter the new implicit words into the sets in the word bag according to the threshold a , and adjust the influence coefficient of existing implicit words according to the threshold e ;
4. $E_n = filter(E_{n-1}, a) \cup adjust(E_{n-1}, e)$;
5. $S_n = filter(S_{n-1}, a) \cup adjust(S_{n-1}, e)$;
6. Adaptive optimization of a and e ;

Return E_n, S_n ;

4 Experiments

4.1 Setting

In this paper, a large number of Chinese corpus with Emoji expressions and special words are required as experimental data. However, there is neither such

a publicly-assessed dataset nor a special word set yet. Therefore, we need to crawl data on the Internet to verify the validity of the model. At the same time, in order to verify the feasibility of the deep memory network with the auxiliary sequences in theory, we implement the sentiment polarity classification with two categories on the IMDB. The evaluation of open source datasets further enhances persuasiveness.

The IMDB two-category dataset is shown in Table 1, and the Chinese data set crawled by the network is shown in Table 2 :

Table 1. The IMDB two-category data set.

Type of set	Positive size	Negative size
Training set	11500	11500
Testing set	1000	1000

Table 2. The Chinese data set crawled.

Type of set	Positive size	Negative size
Testing set	50000	50000
Valid set	5000	5000
Testing set	5000	5000

4.2 Results

We first conduct the evaluation of deep memory network models on IMDB dataset. The learning rate is set to 10^{-3} . One attention layer is set. The batch size is 256. Iterative training is 3 generations. The text word vector is of 50 dimensions. At last about 85% accuracy is obtained on the testing set. The experimental results verify the model’s availability. The accuracy of the experimental results is shown in Fig. 2 and the loss value is shown in Fig. 3.

Second, an evaluation on Chinese dataset with Emoji expressions and some special words crawled on the web, through word2vec is conducted. It is trained on the open source Sougou News data set and 300-dimensional Chinese word vectors are obtained. The learning rate is also set to 10^{-3} . There is also an attention layer. The batch size is 512 and the iterative training is 4 generations. At last, the accuracy rate is about 88% . The accuracy of the deep memory network based on the auxiliary sequence is shown in Fig. 4 and the loss value is shown in Fig. 5.

Emoji expression sets and special word sets are respectively constructed and unified by random vector initialization. Based on the word vector mechanism, a feed-forward network is used to supervise pre-training. Pre-trained 10-dimensional vector representations of Emoji and special words are obtained

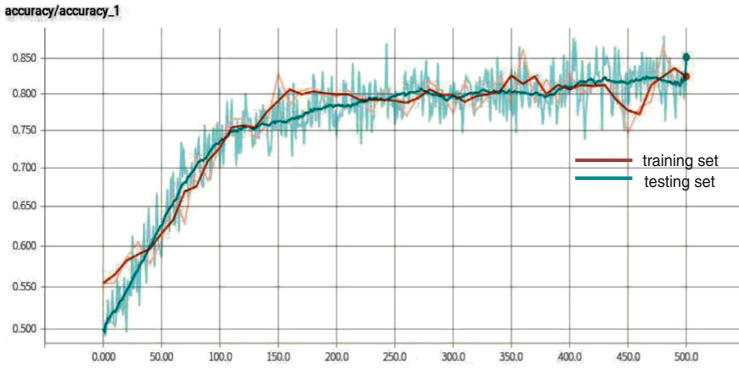


Fig. 2. The accuracy with training steps on IMDB.

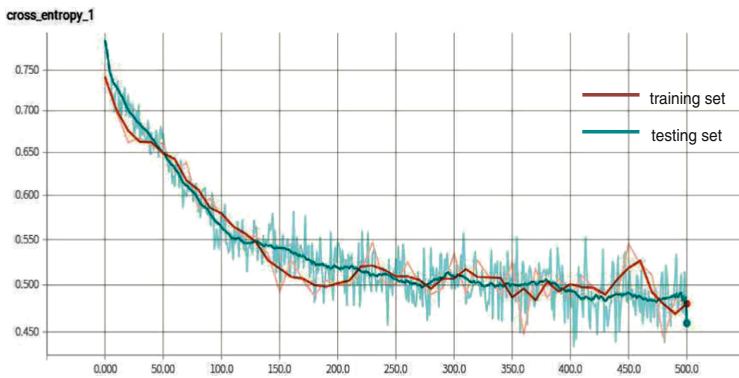


Fig. 3. The loss with training steps on IMDB.

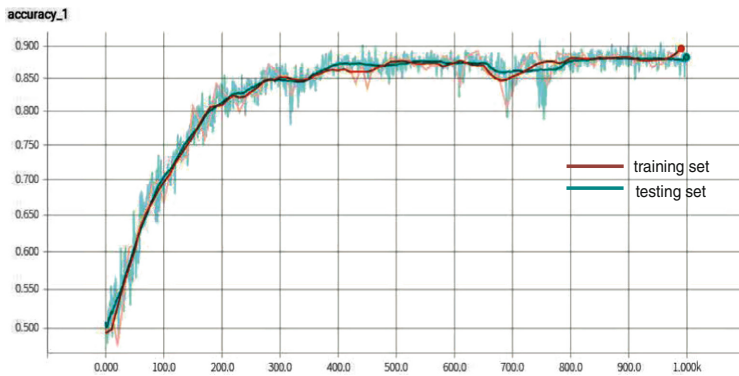


Fig. 4. The accuracy with training steps on Chinese dataset crawled.

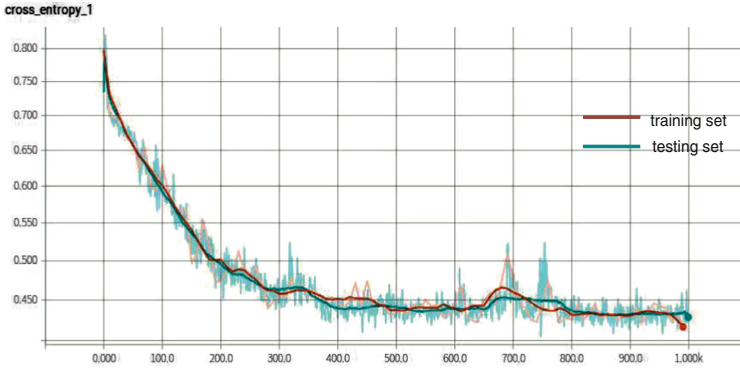


Fig. 5. The loss with training steps on Chinese dataset crawled.

respectively. The 300-dimensional text feature vector, the 10-dimensional Emoji feature vector, and the 10-dimensional special word vector are combined into the final 320-dimensional feature vector for the implicit sentiment classification by the GBDT classifier. The experimental results with accuracy around 96% proves the effectiveness of the model. The confusion matrix is shown in Fig. 6, and the ROC curve is shown in Fig. 7.

4.3 Analysis

With specific examples, we take the sample of Chinese negative sample set for analysis:

Example: I hope all sons and grandsons of the society are like some one which is tall,mighty,honest,cute[tears].

By manual analysis of the above example, considering only the text, it seems to be a positive sentiment. However, the Emoji expression “[tears]” hides the sarcasm, thus the sample sentiment polarity should be negative.

Text Feature Vector. In accordance with the process of the model, the Emoji expression “[Tears]” is temporarily eliminated. A 300-dimensional text feature vector in the corpus is obtained through a deep memory network based on an auxiliary sequence. According to the output vector of the network softmax layer, the prediction is positive class 1, and the classification error occurs.

Emoji Feature Vector. Through the unified construction of Emoji set, a pre-trained “[tears]” Emoji facial 10-dimensional feature vector is obtained. After analysis, the mathematical vector expression is far away from the Euclidean eigenvectors that represent positive emotions such as [haha], which can offset the prediction results of the text vector alone.

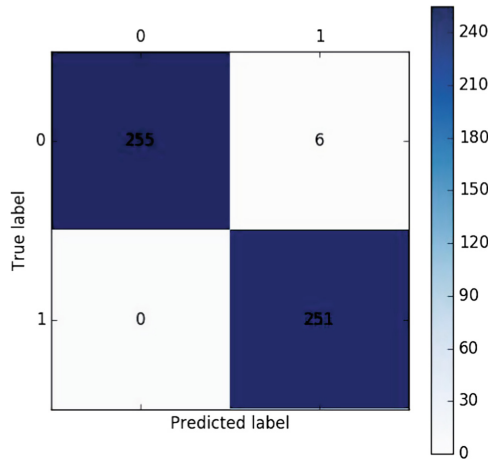


Fig. 6. The confusion matrix.

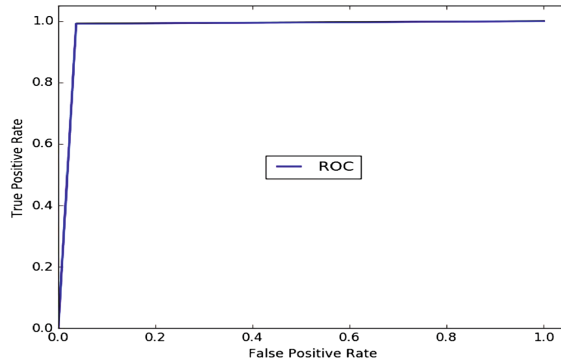


Fig. 7. The ROC curve.

Binary Classifier. Due to the sensitive nature of GBDT for strong and effective features, by inputting the 320-dimensional vectors (include a 10-dimension null special words vector) obtained after the final merging, the example is eventually predicted to be a negative example, thereby realizing the identification of implicit sentiment.

5 Conclusions

Based on the combination of statistics and rules, the recognition of the polarity of the implied sentiment is realized. Combining the attention mechanism and memory network, the auxiliary sequence is used to update the memory in real time, and the text feature vector of the corpus is obtained through deep memory network. Meanwhile, according to the Internet cultural, Emoji set and special

words set are constructed, and they are updated by the online learning method with negative feedback adjustment. Based on word vector mechanism, multilayer perception (MLP) is used to get the feature vectors of Emoji and the special words with implicit sentiment polarity prior knowledge. The text feature vectors and the Emoji, special word feature vectors are combined by extending dimensions to form the final classification feature vectors. Empirical results on IMDB and crawling Chinese data set from Internet verify the feasibility of the theory and the effectiveness of the model.

References

1. Anderson, T.: *The Theory and Practice of Online Learning*. Athabasca University Press, Edmonton (2008)
2. Chikersal, P., Poria, S., Cambria, E.: SeNTU: sentiment analysis of tweets by combining a rule-based classifier with supervised learning. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 647–651 (2015)
3. Cho, K., et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint [arXiv:1406.1078](https://arxiv.org/abs/1406.1078) (2014)
4. Dou, Z.Y.: Capturing user and product information for document level sentiment analysis with deep memory network. In: *Conference on Empirical Methods in Natural Language Processing*, pp. 511–520 (2017)
5. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**(Jul), 2121–2159 (2011)
6. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint [arXiv:1708.00524](https://arxiv.org/abs/1708.00524) (2017)
7. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29**, 1189–1232 (2001)
8. Graves, A., Wayne, G., Danihelka, I.: Neural Turing machines. *Comput. Sci.* (2014)
9. Joshi, A., Sharma, V., Bhattacharyya, P.: Harnessing context incongruity for sarcasm detection. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. vol. 2, pp. 757–762 (2015)
10. Khan, A.Z., Atique, M., Thakare, V.: Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *Int. J. Electron. Commun. Soft Comput. Sci. Eng. (IJECSCE)* 89 (2015)
11. Kumar, A., et al.: Ask me anything: dynamic memory networks for natural language processing, pp. 1378–1387 (2015)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436 (2015)
13. Salton, G., Fox, E.A., Wu, H.: Extended Boolean information retrieval. *Commun. ACM* **26**(11), 1022–1036 (1983)
14. Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R.: End-to-end memory networks. *Comput. Sci.* (2015)
15. Tang, D., Qin, B., Liu, T.: Aspect level sentiment classification with deep memory network, pp. 214–224 (2016)
16. Weston, J., Chopra, S., Bordes, A.: Memory networks. Eprint Arxiv (2014)

17. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.: Hierarchical attention networks for document classification. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1480–1489 (2017)
18. Zhang, L., Ghosh, R., Dekhil, M., Hsu, M., Liu, B.: Combining lexicon-based and learning-based methods for twitter sentiment analysis. Hp Laboratories Technical Report (2011)



Intelligent Traffic Light System for High Priority Vehicles

Guiduan Li, Guozhi Song, and Wen Li^(✉)

School of Computer Science and Technology,
Tianjin Polytechnic University, Tianjin 300387, China
intwentytwelve@163.com

Abstract. With the development of the globalization, the living standard has been improved. The increased number of vehicles on road made the method of controlling the traffic light which is traditional and empirical with poor efficiency. The default traffic light system can not satisfy peoples travel demand especially on the congested intersection. The intelligent traffic light system can adapt to the flow at the intersection and change the Traffic Light Duration Cycle (TLDC) in order to reduce travel time of all vehicles. Moreover, there are High Priority Vehicles (HPV) on road, designed to reach the destination on time. So they should be given privileges to avoid traffic jams. The proposed work, based on the priority of vehicles, aims at providing an intelligent traffic light system which the HPV can send request to after be loaded at junction. According to the highest priority, System would turn traffic light green to clear the Road Segment (RS) for saving travel time of HPV. The system is tested on Simulation of Urban Mobility (SUMO) and use the Traffic Control Interface (TraCI) of Python. The results show the effectiveness of the intelligent traffic light system. It may has significant theoretical as well as practical value for Intelligent Transportation System (ITS) in the future.

Keywords: Intelligent traffic light system · Traffic Light Duration Cycle · High Priority Vehicles · Simulation of Urban Mobility · Traffic Control Interface

1 Introduction

With the rapid development of Internet of things (IOT) technology and industry, many new concepts emerge into our lives. It is convenient for modern people to lead quality lives and improve the safety factor. Connected car [1]—It is a car equipped with Internet access, and usually also has a wireless LAN. It allows cars and other equipment inside and outside the car to interface the Internet and realize device sharing and data sharing. Typically, cars are also equipped

Supported by National Natural Science Foundation of China (No. 61972456).

© Springer Nature Singapore Pte Ltd. 2019
S. Guo et al. (Eds.): CWSN 2019, CCIS 1101, pp. 212–223, 2019.
https://doi.org/10.1007/978-981-15-1785-3_16

with special technologies such as the Internet or WLAN to facilitate drivers manipulation.

Nowadays, the problem of urban traffic congestion has become one of the most difficult reasons for people to travel. On the one hand, those commuters who spend much time on waiting the traffic jams would be anxious to drive cars, thereby leading to traffic accidents. On the other hand, the HPV should have priority to pass the junction rapidly as possible as they can. HPV are the vehicles used in public emergencies which include ambulances, fire engines, patrol wagon and so on [2]. So Traffic lights installed at various junctions are the most common and effective means of regulating vehicles.

In our country, almost all cities adopt the traditional traffic control system. The traffic lights are usually controlled by timing which means the Traffic Light Duration Cycle (TLDC) is fixed, so the actual traffic flow can not be identified and optimized. That is to say it can not adapt to the uncertainty and randomness of traffic flow, which often results in the waste of traffic resources and the congestion of roads. What is more, although some cities have studied and brought in some advanced traffic light management systems, the rate of traffic accident remains high due to the shortage of infrastructures and other reasons. The intelligent traffic control system can improve the transportation efficiency effectively as well as safety factor without hardware changes.

In this work, we first introduce the contribution of researchers to the intelligent traffic light system. Then we focus on reducing the travel time of cars as an approach to improve efficiency of traffic light system. So we design an intelligent traffic light system which can change the traffic light duration cycle according to the flow of the Road Segment (RS) and make a comparison with the normal mode about the total travel time of HPV. Subsequently, we classify the priority of HPV, the higher priority of HPV, the quicker to pass the intersection. In the end, we conclude that proper traffic light cycle time and priority work would increase the mobility of HPV by taking intelligent decisions.

2 Related Work

Many researchers have contributed to improve the traffic light scheduling, making a proper traffic light synchronization while making efforts to optimize the time that all the vehicles wait at the intersection.

In the aspect of ITS, M Tubaishat et al. proposed a traffic light control system based on WSN [3] in order to reduce the waiting time of vehicles. M Collotta et al. designed a traffic light management system combining fuzzy logic and WSN [4]. A Leyre et al. has evaluated the deployment challenges of signal integration in wireless sensor networks [5].

For intelligent traffic light, optimization simulation method based on PSO (Particle Swarm Optimization) algorithm [6] can maximize the number of traffic passes or minimize vehicle travel time in the shortest simulation time. Shortest path algorithm model [7] finds the shortest path for HPV to reach the destination. It calculates the distance between HPV and destination and use Dijkstra

algorithm to find real time dynamic shortest path on the spot. Genetic algorithm [8] can reduce the length of vehicle queue in ITS. They conclude that it is useful to use evolution strategies to solve this type of problems.

The biggest difference between intelligent traffic lights and traditional traffic lights is that they can automatically adjust the duration of traffic lights according to the current traffic flow and then reduce the average travel time of vehicles. Therefore, in order to achieve the best control effect of traffic lights, it is indispensable to explore the suitable periodic control algorithm [9]. However, if we used real cars on road in the primary stage of the algorithm control test, it would not only waste a lot of manpower and material resources, but also cause vehicle collision and affect the traffic. Therefore, the method of virtual simulation meets the requirements of the test well.

3 Proposed System

3.1 Variable Periodic Traffic Light Model

The simulation of intersection to be built uses induction loops and defined interface in SUMO to collect data. We need to find the variables to be monitored according to the control principle of traffic light.

Traffic light has four phases:

1. Green light of horizontal direction and red light of vertical direction.
2. Yellow light of horizontal direction and red light of vertical direction.
3. Green light of vertical direction and red light of horizontal direction.
4. Yellow light of vertical direction and red light of horizontal direction.

The introduction about control principle of traffic light is as follows: Monitoring the traffic flow in each phase real time and using periodic calculation formula [9] to get the period of traffic light under its corresponding traffic flow. Then the green light duration time (also called effective green time) of each phase is assigned according to the green signal ratio allocation formula.

In the book Automatic Control of Road Traffic, the author proposes that the optimal cycle time [10] of minimum delay of vehicles when the traffic flow is stable and the time of vehicle arrival at the intersection is random is:

$$C_o = \frac{1.5L + 5}{1 - Y} \quad (1)$$

where L represents total lost time within a cycle (start delay time and end lag time), Y is the sum of the maximum saturation of the critical phases of the period.

$$L = \sum_{i=1}^n (l_i + I_i) \quad (2)$$

where l_i is the lost time of vehicles and I_i is the green light interval time.

$$Y = \sum_{i=1}^n \max(y_i, y_i' \dots) \quad (3)$$

where y_i is the maximum saturation (flow ratio) of phase i .

$$y_i = \frac{q_i}{s_i} \tag{4}$$

where q_i represent actual arrival flow of phase i , s_i is saturation flow of phase i . Saturation of different directions [11] is shown in Table 1.

Table 1. Basic saturated flow of entrance lane.

Lane	$S_{bi}(pch/h)$
Straight	1550–1750
Left-turn	1350–1550
Right-turn	1450–1650
Straight and right-turn	1150–1350
Straight and left-turn	1150–1350

Then allocate the green light time according to the Green Split [10]:

$$G_e = C_0 - L \tag{5}$$

where G_e represents the effective green time in TLDC.

$$g_{e1} = G_e * \frac{\max(y_1, y_1')}{Y} \tag{6}$$

$$g_{e2} = G_e * \frac{\max(y_2, y_2')}{Y} \tag{7}$$

where g_{e1} and g_{e2} are the effective green time in that phase.

3.2 Priority Pass Model of HPV

Then the priority work concentrate on reducing traffic congestion for HPV. It can realize an interactive system which the HPV driver can send request to the centralized traffic control system and system calculates the priority of each RS. Finally, the system can turn the traffic light green for the RS with the highest priority.

When there are no HPV, the traffic light will run in the default manner. Suppose that there is a high priority vehicle in the system, system will turn the traffic light green for that RS. If the two or more HPV passed the induction loop, system would receive the requirements sent by HPV drivers and calculate the priority of each RS. In this way, the RS which has the highest priority over others would turn traffic light green.

The significant formula for calculating priority of a RS [2] is as follows:

$$P_{val}(RS) = Req_num * W_{Req_num} + \frac{W_{Min_dist}}{Min_dist} + Max_wait * W_{Max_wait} + Amb_val * W_{Amb_val} \quad (8)$$

$P_{val}(RS)$ represents the priority of a RS at an intersection, Req_num is number of HPV requests from a RS, Min_dist is distance of closest HPV on a RS from a traffic light at an intersection, Max_wait is maximum of HPVs waiting time on a RS of an intersection. Amb_val indicates the emergency level of the HPV, ranging from 1 to 10. The higher the Amb_val , the more urgent of the HPV.

W_{Req_num} , W_{Min_dist} , W_{Max_wait} , W_{Amb_val} are the weights of Req_num , Min_dist , Max_wait and Amb_val respectively.

The weights have a close relationship with circumstance of RS. The value of the weights depends on the design of the road map, traffic intensity, population and so on. In our experiment, $W_{Req_num} = 1$, $W_{Min_dist} = 100$, $W_{Max_wait} = 0.1$, $W_{Amb_val} = 1$ [2].

Case I: When there are no HPV on RS, $Req_num = Min_dist = Max_wait = Amb_val = 0$. So the traffic light will run in the default manner.

Case II: When there are HPV on RS, $P_{val}(RS)$ is not zero. And traffic light turns green for RS which has the highest priority.

4 Implementation

We have connected this development with the widely recognized traffic simulator SUMO [12]- it is an open source, microscopic, multi-modal traffic simulation. It allows to simulate how a given traffic demand which consists of single vehicles moves through a given road network and address a large set of traffic management topics.

The road network includes 9 nodes and 16 edges. The communication between road net and SUMO is TraCI (Traffic Control Interface) tool [13]. Giving access to a running road traffic simulation, it allows to retrieve values of simulated objects and to manipulate their behaviour “on-line”. In the mean while, it is based on a client/server architecture that allows to control simulations through the script files. The intelligent traffic light system is implemented by Python.

4.1 Establishment of an Intersection that Can Automatically Collect Data

According to the formula of optimal cycle time, traffic variables to be detected of intelligent traffic light are: traffic flow of each RS, start delay time, end lag time. We need to monitor the travel time of HPV as well in order to comparing control effects.

Specifically, traffic flow represents the number of vehicles leaving the current intersection per unit time, start delay time means the time between the end of the

red light and the start of the first waiting car. End lag time is the time between the end of the green light and the arrival of the last car at the intersection.

We build up an intersection and each road has two directions. So as to achieve the intelligent control of traffic light, intersection has to own the ability of real-time data acquisition and control. So we bring in the induction loop [14] of SUMO which can monitor the traffic variables directly or indirectly in Fig. 1.

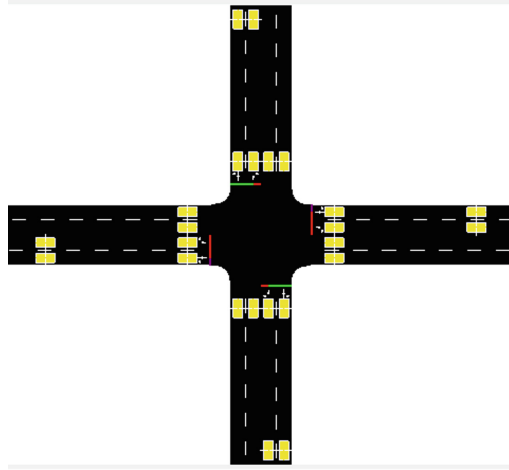


Fig. 1. Automatically collect data intersection. (Color figure online)

In order to get precise start delay time and end lag time, we bring in a new model which includes three induction loops in each direction. One is put at the entrance of intersection and another is put at exit of intersection. The last one is a little far from entrance of intersection.

When a car stop on an induction loop at the entrance of intersection, system would calculate the time difference between the end of the red light and the start of the car. And it is called start delay time.

As for end lag time, when the induction loops at entrance and exit of intersection detect the difference of number of cars is 1, that means the arrival of last car at the intersection. When the induction loops at entrance and exit of intersection detect the difference of number of cars is 0, that means the traffic light is going to turn red. The time difference between 2 states is end lag time.

The model is being simulated under three traffic conditions i.e. Low traffic (104 vehicles and 15 HPV), Moderate traffic (237 vehicles and 15 HPV), and High traffic (385 vehicles and 15 HPV). HPV's destination is set at (10, 510). HPV followed a definite path between source and destination.

Similarly, in order to be more realistic, we gave up the cars model which have fixed route and starting time. All the vehicles are randomly generated and the routes of vehicles are random as well. The default vehicle is a modification of the

model defined by Stefan Krauß [15]: Let vehicles drive as fast as possibly while maintaining perfect safety. Normal vehicles are yellow in SUMO and HPV are red.

For implementation of this model we take two modes into consideration: Normal Mode and Intelligent Mode. Normal Mode is the traffic light controlled by the default manner while Intelligent Mode use the variable periodic algorithm. Time taken by HPV is noted in both modes. Finally, We compared the total travel time of HPV.

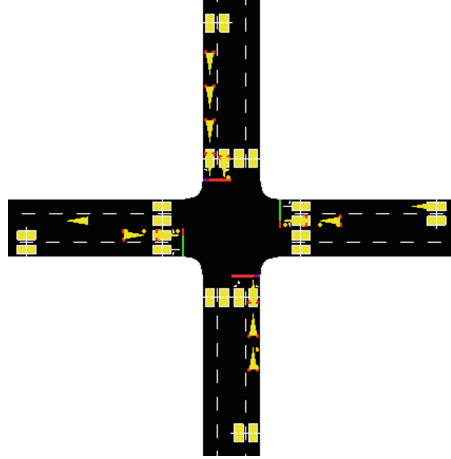


Fig. 2. No HPV on RS. (Color figure online)

4.2 Working of Priority Model

For implementation of this model we take three modes into consideration: Normal Mode, Proposed Mode and Priority Mode. Normal Mode is the traffic light controlled by the default manner. Proposed Mode would turn the traffic light green as soon as the HPV pass the induction loop from every phase on RS. And Priority Mode calculate the priority of each RS and turn traffic light green for RS which has the highest priority.

So as to saving HPVs travel time, we bring in a theory of congestion: when there are no HPV on RS and no more than 5 vehicles (35 m from the intersection) at the intersection, the traffic light run in the default manner in Figs. 2 and 3 shows there are 5 vehicles at the intersection from the north, we consider that as the situation of traffic congestion because the distant induction loop can perceive a static car on it. So we set the traffic light green when there are 5 or more vehicles waiting at the intersection. This method can effectively reduce probability of congestion and save travel time of HPV which can let HPV arrive at destination as soon as possible.

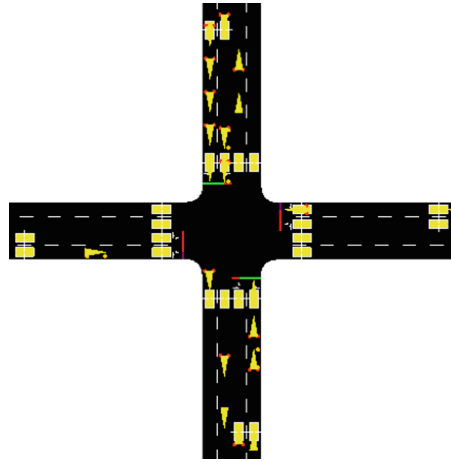


Fig. 3. Traffic congestion at the intersection. (Color figure online)

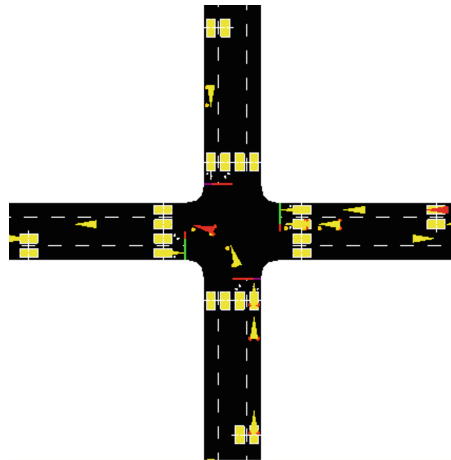


Fig. 4. One high priority vehicle in Proposed Mode. (Color figure online)

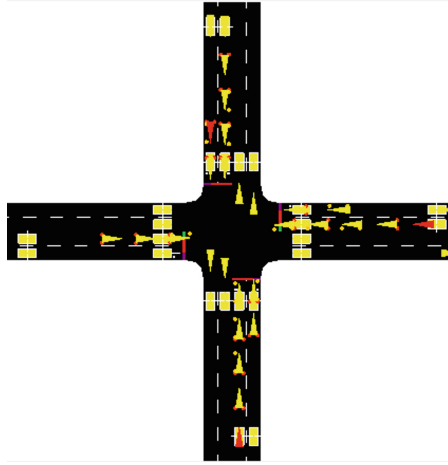


Fig. 5. Two HPV in Proposed Mode. (Color figure online)

Proposed Mode. In the Proposed Mode, system would turn the traffic light green when there is 1 high priority vehicle passes induction loop in Fig. 4. When there are 2 induction loops monitor the HPV at the same time, system would use the FCFS (First Come First Served) to regulate the traffic light. Figure 5 shows there are 2 HPV from the east and south, Proposed Mode set the southern RS green.

Priority Mode. In the Priority Mode, system would turn the traffic light green when there is 1 high priority vehicle passes induction loop in Fig. 6. When there are 2 induction loops monitor the HPV at the same time, system calculate the priority of each RS and turn traffic light green for RS which has the highest priority. Figure 7 shows there are 2 HPV from the east and south, the eastern priority of RS is higher than the southern. So the Priority Mode set the eastern RS green.

5 Result

5.1 Variable Periodic Traffic Light Model

In the variable periodic traffic light model, we compare the total travel time of high priority vehicles under three traffic densities between two modes. The calculated optimal cycle time of different traffic densities is show in the Table 2.

As shown in the Fig. 8, in the Low density traffic intersection, the Normal Mode HPV travel time is 211 s more than the Intelligent Mode. In the moderate density traffic intersection, the Normal Mode HPV travel time is 852 s more than the Intelligent Mode. In the high density traffic intersection, the Normal Mode HPV travel 555 s longer than the Intelligent Mode.

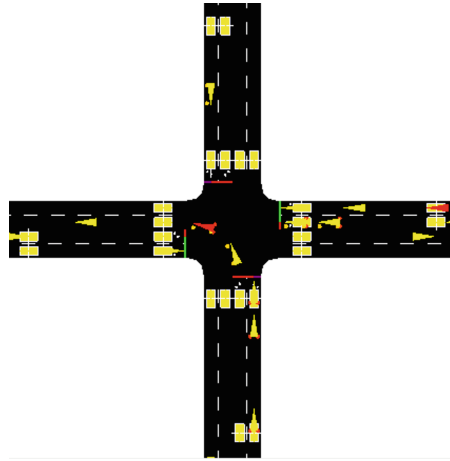


Fig. 6. One high priority vehicle in Priority Mode. (Color figure online)

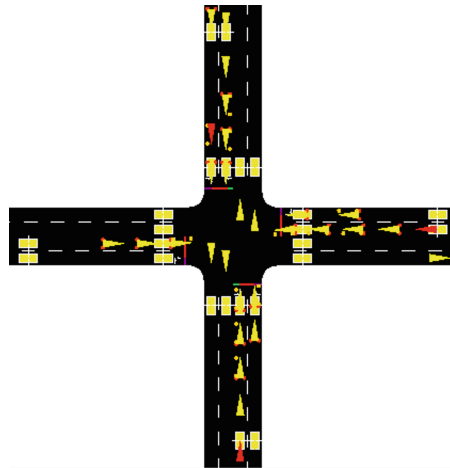


Fig. 7. Two HPV in Priority Mode. (Color figure online)

Table 2. Optimal cycle time of different traffic densities.

Traffic conditions	Default duration cycle time (s)	Optimal cycle time (s)	Effective green time in first phase (s)	Effective green time in second phase (s)
Low	90	17	6.24	4.76
Moderate	90	46	22	18
High	90	53	25	18

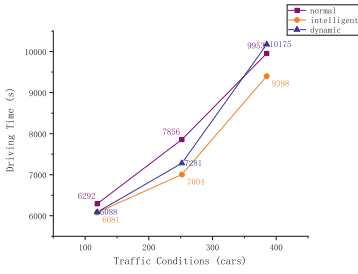


Fig. 8. Total travel time of HPV under three traffic densities between two modes.

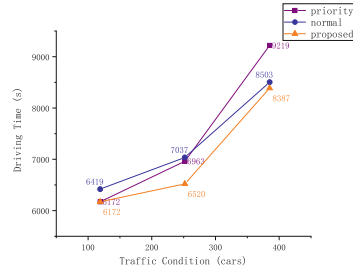


Fig. 9. Total travel time of HPV under three traffic densities among three modes.

In the Priority pass model, we compare the total travel time of high priority vehicles under three traffic densities among three modes.

As shown in the Fig. 9, in the Low density traffic intersection, the Proposed Mode HPV travel time is the same as Priority Mode with 6172 s. Normal Modes is 247s more than the other modes. In the moderate density traffic intersection, the Proposed Mode HPV travel time is the least with 6520 s while Normal Mode is 7037 s. The HPV travel time of Priority Mode is 74s less than Normal Mode. In the high density traffic intersection, the situation is quite different from the other two traffic conditions. Priority Mode waste too much time with 9219s, the travel time of Normal Mode and Proposed Mode are close, with 8503 s and 8387 s respectively.

6 Conclusion

This paper presents two models for traffic control using a simulator SUMO and TraCI as an interface to achieve real-time monitoring traffic conditions. The development is focus on reducing the traffic congestion which is the most relevant challenge in modern cities. Specifically, the main goal is to reduce the total travel time of HPV.

In this work, We designed and implemented a variable periodic traffic light system. It can adapt TLDC to the traffic density. After analyzing the results, it is an effective method of relieving the traffic jams and reducing the service time of HPV. Particularly, in the moderate density traffic intersection, the effect is the most obvious with 852s declination. Then we brought in a priority model, aimed at letting the HPV have the priority to pass the junction. By using this model, we can conclude that in low/moderate traffic condition, it can reduce the impacts of city traffic. But in high traffic condition, the Proposed Mode is better than Priority Mode. The results are in accordance with the practice and there are some limits of traffic density for using them.

However, it can be easily extended to artificial Intelligence techniques, particularly unsupervised learning, to create a more robust system for real applications in the future.

References

1. Wu, S.: Connected car: a subject worthy of concern. *Chin. Telecoms Ind.* **6**(8), 17–19 (2010)
2. Dang, D., Tanwar, J., Masood, S.: A smart traffic solution for High Priority Vehicles. In: 2015 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, pp. 466–470 (2015). <https://doi.org/10.1109/NGCT.2015.7375162>
3. Tubaishat, M., Qi, Q., Shang, Y., Shi, H.: Wireless sensor-based traffic light control. In: 2008 5th IEEE Consumer Communications and Networking Conference, Las Vegas, pp. 702–706 (2008). <https://doi.org/10.1109/ccnc08.2007.161>
4. Collotta, M., Bello, L.L., Pau, G.: A novel approach for dynamic traffic lights management based on wireless sensor networks and multiple fuzzy logic controllers. *Expert Syst. Appl.* **42**(13), 5403–5415 (2015). <https://doi.org/10.1016/j.eswa.2015.02.011>. ISSN 0957–4174
5. Azpilicueta, L., et al.: Evaluation of deployment challenges of wireless sensor networks at signalized intersections. *Sensors* **16**(7), 1140 (2016). <https://doi.org/10.3390/s16071140>
6. Panovski, D., Zaharia, T.: Simulation-based vehicular traffic lights optimization. In: 2016 12th International Conference on Signal-Image Technology Internet-Based Systems SITIS, Naples, pp. 258–265 (2016). <https://doi.org/10.1109/SITIS.2016.49>
7. Arunmozhi, P., William, P.J.: Automatic ambulance rescue system using shortest path finding algorithm. **3**(5), 635–638 (2014)
8. Teo, K.T.K., Kow, W.Y., Chin, Y.K.: Optimization of traffic flow within an urban traffic light intersection with genetic algorithm. In: 2010 Second International Conference on Computational Intelligence, Modelling and Simulation, Tuban, pp. 172–177 (2010). <https://doi.org/10.1109/CIMSiM.2010.95>
9. Liu, S., Li, X., et al.: Building simulation environment with SUMO to support intelligent traffic light. *Inf. Traffic* (10), 36–37 (2016)
10. Yan, R.: Study on the Method of Determining the Optimal Period of Traffic Light and its Simulation. *Traffic Information Engineering and Control*. Northeast Forestry University (2010)
11. Duan, L.: *Road Traffic Automatic Control*. Chinese People's Public Security University Press, Beijing (1991)
12. Seah, W.K.G.: *The International Journal on Advances in Systems and Measurements is Published by IARIA* (2009)
13. Wegener, A., Raya, M., et al.: TraCI: an interface for coupling road traffic and network simulators. In: *Proceedings of the 11th Communications and Networking Simulation Symposium*, pp. 155–163. ACM, New York (2008). <https://doi.org/10.1145/1400713.1400740>
14. Cruz-Piris, L., Rivera, D., Mars-Maestre, I., de la Hoz, E., Fernandez, S.: Intelligent traffic light management using multi-behavioral agents (2017)
15. SUMO Homepage. <http://sumo.sourceforge.net/userdoc/>. Accessed 27 Apr 2019



Personalized Recommendation Based on Tag Semantics in the Heterogeneous Information Network

Bin Yan^{1,2,3}, Lichen Zhang^{1,2,3}(✉), Longjiang Guo^{1,2,3}, Meirei Ren^{1,2,3},
and Ana Wang^{1,2,3}

¹ Key Laboratory of Modern Teaching Technology,
Ministry of Education, Xian 710062, China
zhanglichen@snnu.edu.cn

² Engineering Laboratory of Teaching Information Technology of Shaanxi Province,
Xian 710119, China

³ School of Computer Science, Shaanxi Normal University, Xian 710119, China

Abstract. Heterogeneous information network (HIN) is widely used in recommendation system because of its superiority in complex information modeling. However, the existing HIN-based methods ignore two issues. First, low-quality information may cause users to be dissatisfied with the recommendation results. Secondly, HIN-based recommendations are difficult to predict the user's attitude toward the item. Therefore, this paper proposes two improvement strategies: (1) propose a semantic information filtering strategy to filter low-quality information and improve recommendation efficiency; (2) integrate tag information into HIN-based recommendation system to achieve personalization recommend. This paper verifies the validity of the proposed model on two real data sets.

Keywords: Recommender system · Heterogeneous information network · Tag semantic

1 Introduction

In order to solve the problem of information overloads with the information network era, a variety of recommendation algorithms are proposed. For example, the classic collaborative filtering algorithm [1–3] uses the similarity between users or items for recommendations. The heterogeneous information network (HIN) is used as a modelling method to deal with the complexity and heterogeneity of information [4, 5]. Some methods prove that tag semantic information can effectively analyze the user's personalized features [6, 7].

The main work of this paper consists of two aspects. Firstly, the HIN information is rated and an information filtering strategy is proposed to reduce the negative impact of low-quality information on recommendation. Secondly, the HIN semantic The main work of this paper consists of two aspects. Firstly, the

HIN information is rated and an information filtering strategy is proposed to reduce the negative impact of low-quality information on recommendation. Secondly, the HIN semantic.

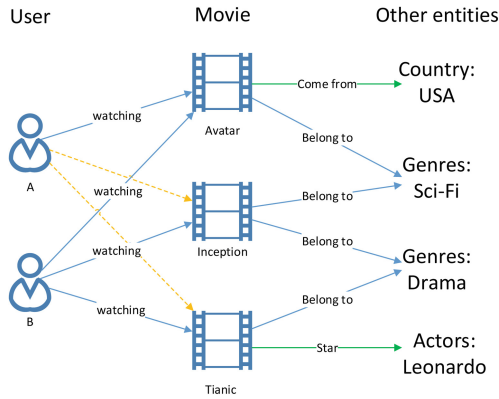


Fig. 1. Movie heterogeneous information network.

Most of the existing HIN-based recommendation methods use the semantic correlation based on the meta-path to recommend [8–10]. In other words, these methods use the meta-path to obtain the relevant semantic information about the recommendation. As shown in the figure of a movie HIN, different entity types (users and movies, movies and genres) have different semantic relationship types (watching, belonging to). The movie inception and the movie avatar watched by user A belong to the same genre, so there may be a potential relationship between user A and inception.

On this basis, many improvement methods are proposed. For example, Zhao et al. [11] use the meta-graph instead of meta-path fusion, Dong et al. and Fu et al. [12, 13] use network embedding for information mining, Zhu et al. [14] merge multiple HINs to alleviate cold-start and data sparsity problems. In addition, various auxiliary data is also widely used, for example, Feng et al. [7] build HIN with tag information, or Hu et al. [15] add context information of the meta-path.

However, the meta-path can find potential semantic relationships between users and items in HIN, but does not reflect user attitude towards the item. Although the tag semantics can reflect the user’s personal preferences, but lack of explanation for the recommendations, this paper considers to combine the two for recommendation. In addition, the above method does not consider the influence of information quality on recommendation, while low-quality information may cause users to be dissatisfied with the recommendation results. This paper considers the measurement of information quality, and then proposes a filtering strategy. Our main contributions are as follows.

- (1) This paper proposes a method for filtering low-quality information in HIN.
- (2) This paper integrates tag information into the HIN-based recommendation system.
- (3) This paper validates valid items on real data sets.

The rest of this paper is organized as follows Sect. 2 introduces the background and preliminary knowledge, Sect. 3 mainly introduces our recommendation model, Sect. 4 reports the experimental results, Sect. 5 concludes our work.

2 Background and Preliminaries

2.1 Heterogeneous Information Network

A heterogeneous information network (HIN) [9] is defined as a directed graph $G = (V, E)$, with an entity type mapping function $\phi : V \rightarrow \mathcal{A}$ and a link type mapping function $\varphi : E \rightarrow \mathcal{R}$, \mathcal{A} denotes the entity typeset and \mathcal{R} denote the link typeset, where $|\mathcal{A}| + |\mathcal{R}| > 1$.

Figure 2 shows partial network schemas of heterogeneous movie information network. Users and movies are different types of entities in the network, and links between movies, users and countries represent different types of relationships. For instance, the relationship between movies and users is “watching”, while the relationship between movies and countries is “belonging”. The two types are different.

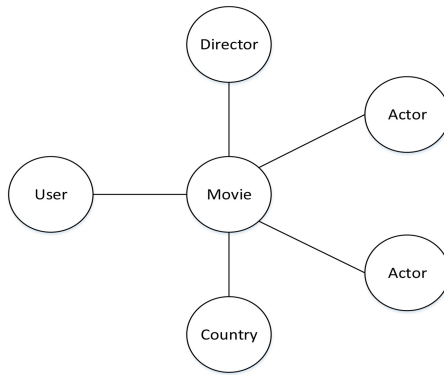


Fig. 2. Movie HIN schemas.

2.2 Meta-path

A meta-path can be expressed as $A_0 \xrightarrow{\mathcal{R}_1} A_1 \xrightarrow{\mathcal{R}_2} \dots \xrightarrow{\mathcal{R}_n} A_n$ in HIN, we define $\mathcal{R} = \mathcal{R}_1 \circ \mathcal{R}_2 \circ \dots \circ \mathcal{R}_n$ represents the composite relation that exists between

entity types A_1 and A_n ($A_i \in \mathcal{A}, R_j \in \mathcal{R}$), and \circ represents the composition operator [15,17].

As shown in Fig. 2, users and movies can be connected through the “User-Movie-genres-Movie” (UMAM) path represents the actor appearing in two movies, where the “genres” can be replaced with “director”, “actors”, “country” and so on. Through the meta-path, we can find the potential relationship between the user and the item, which can enhance the interaction between the user and the item. At the same time, we use pathsim [9] to calculate the similarity score of the two entities on the meta-path.

3 The Proposed Model

In this section, we propose a personalized recommendation based on tag semantics in the heterogeneous information network - PTRS. Algorithm 1 represents the three phases of the PTRS model - user clustering, filtering and prediction.

Algorithm 1. PTRS Scheme

Input:

set of user U , and the set of user u_i tags $T_u^{(i)}$
 set of movie V , and the set of movie v_j tags $T_v^{(j)}$
 set of meta-path P

Output:

Recommendation for requests

- 1: **Procedure** PTRS
 - 2: **Procedure** User cluster phase
 - 3: Use K-means algorithms to cluster users
 - 4: **End procedure**
 - 5: **Procedure** Filter phase
 - 6: Categorize and rate semantic information
 - 7: Filter
 - 8: **End procedure**
 - 9: **Procedure** Prediction phase
 - 10: predicting unrated items for requests
 - 11: Making recommendation
 - 12: **End procedure**
 - 13: **End procedure**
-

3.1 User Cluster Phase

In the actual situation, the data is very sparse, and the single meta-path filtering effect is not good. Therefore, we first cluster users and then perform meta-path filtering on user sub-groups to alleviate data sparsity.

We use the Cosine Similarity to calculate the similarity between users, use the K-Means algorithm cluster the user set U into $\{UC_1, UC_2, \dots, UC_N\}$.

3.2 Filter Phase

In HIN, we can express the recommended process as a process of finding potential semantic relationships, and the meta-path is an important tool to implement the above process. For example, in Fig. 1, we recommend the movie inception for User A through the “UMGM” meta-path, which contains semantic information such as “Users Watched Movie Avatar”, “Avatar and Inception belong to Sci-Fi movies”, and movie genre play a key role (“Sci-Fi” is one of the key words of the meta path). If we analyze the user’s tag information and get the user does not like Sci-Fi movies, then it should not be established for user A to recommend Inception, and the semantic information related to “Sci-Fi” in the meta-path is low quality information. To solve this problem, our filtering strategy is to divide the original meta-path into multiple sub-meta-paths by keywords (For example, in the meta path UMGM, the keywords include “Sci-Fi”, “Drama”, “Romance”, “Musical” etc., We group the keywords “Sci-Fi” and “Drama” into one group, and the other keywords are grouped into one group, so that the UMGM meta path can be divided into two sub-meta paths), and then use the tag information of users to distinguish low-quality meta-paths.

In user subgroup UC_k , Meta-path set is $P = \{p_1, \dots, p_q, \dots, p_L\}$. We can calculate the similarity between the user u_i ’s tag t_u and meta-path p_q ’s keyword τ_q as follows.

$$sim(t_u, \tau_q) = \frac{2 \cdot depth(lso(t_u, \tau_q))}{len(t_u, \tau_q) + 2 \cdot depth(lso(t_u, \tau_q))} \quad (1)$$

where $depth(t_u)$ represents the depth of t_u in the WordNet relational tree, $len(t_u, \tau_q)$ is the shortest distance between two words, and $lso(t_u, \tau_q)$ represents the common parent node of two words in the relational tree.

We can further calculate the user u_i rating of the meta-path p_q as follows.

$$S(u_i, p_q) = \frac{\sum_{t_i \in T_u^{(i)}} \sum_{\tau_j \in \Gamma_q} sim(t_i, \tau_j) \times \bar{r}_{t_i}^{(u_i)}}{|T_u^{(i)}|} \quad (2)$$

The user subgroup UC_k rates the meta-path as follows.

$$S(UC_k, p_q) = \frac{1}{|UC_k|} \sum_{u_i \in UC_k} S(u_i, p_q) \quad (3)$$

where $\bar{r}_{t_i}^{(u_i)}$ denotes the average score of the movies marked by tag t_i , $T_u^{(i)}$ denotes the set of tags for the user u_i , Γ_q denotes the set of tags for the meta-path p_q . We filter out every rating of meta-path which is less than 1.

3.3 Prediction Phase

As we discussed earlier, the meta-path in HIN can find potential semantic relationships between users and items, but does not reflect user attitudes toward items. Although tag semantics can reflect a user's personal preferences, there is a lack of explanation for recommendations. For example, if a user likes a science fiction movie, the HIN-based recommendation system can recommend the Transformers series to him, but we don't know which movie the user prefers. Combined with the semantics of the tag, we found that the tags of the Transformers series are different, and Transformers 1 is more in line with the user's taste, so it should be given priority.

We can get L matrix $\{\hat{R}_k^{(1)}, \hat{R}_k^{(2)}, \dots, \hat{R}_k^{(L)}\}$ through L element paths of user subgroup UC_k . Where denotes the q -th score matrix in UC_k . Then we use the Low-Rank Matrix Factorization to process the matrix, and get two low-rank latent representation matrices (U and V) of users and items as follow.

$$\begin{aligned} (\hat{U}_k^{(q)}, \hat{V}_k^{(q)}) &= \arg \min_{U, V} \left(\frac{1}{2} \|\hat{R}_k^{(q)} - UV^T\|_F^2 + \frac{\lambda}{2} \|U\|_F^2 + \frac{\lambda}{2} \|V\|_F^2 \right) \\ \text{s.t. } & U \geq 0, V \geq 0 \end{aligned} \quad (4)$$

where $\frac{\lambda}{2} (\|U\|_F^2 + \|V\|_F^2)$ is the quadratic regularization term that avoids over fitting, λ is a regularization parameter that determines the importance of the quadratic regularization term, $\|*\|_F$ is the *Frobenius norm*.

We can further get L pairs of implicit feature matrix of users and items $(\hat{U}^{(1)}, \hat{V}^{(1)}), \dots, (\hat{U}^{(L)}, \hat{V}^{(L)})$. Where the matrix U represents the degree of preference of users for different features, and the matrix V represents the degree of importance of different features in items. We define the calculation function of user u_i 's preferences for item based on the meta-path.

$$r^p(u_i, v_j) = \sum_{k=1}^N \text{Cos Sim}(UC_k, u_i) \sum_{q=1}^L \omega_k^{(q)} \cdot \hat{U}_i^{(q)} \hat{V}_j^{(q)} \quad (5)$$

We calculate the calculation function of user u_i 's preferences for movies by tag semantics.

$$r^T(u_i, v_j) = \frac{\sum_{t_u \in T_u^{(i)}} \sum_{t_v \in T_v^{(j)}} \text{sim}(t_u, t_v) \cdot \bar{r}_{t_u}^{(i)}}{|T_u^{(i)}| |T_v^{(j)}|} \quad (6)$$

Then, user u_i 's preference to item \mathcal{V}_i as follow.

$$r^* = (1 - \alpha)r^p + \alpha r^t \quad (7)$$

where α represents the weight of tag semantics in the recommendation. In addition, we can also predict the user u_i 's rating function of the movie \mathcal{V}_j .

$$s(u_i, v_j) = \frac{\bar{r}(u_i) \cdot r^*(u_i, v_j)}{\bar{r}_{u_i}^*} \quad (8)$$

where $\bar{r}_{(u_i)}$ represents the average of all user’s scores, and $\bar{r}_{u_i}^*$ represents the average of user’s preferences for other movies.

3.4 The Model Optimization

For the user subgroup UC_k , the weight of the multi-segment path fusion will have a significant impact on the recommendation result. We represent the weight vector as $\omega = \{\omega^{(1)}, \dots, \omega^{(q)}, \dots, \omega^{(L)}\}$, $\omega^{(q)}$ represents the weight of the path q , then the final score prediction matrix is as follows.

$$\hat{R} = \sum_{q=1}^{|P|} \omega_k^{(q)} \times \hat{R}^{(q)} \quad (9)$$

where $\omega_k^{(q)}$ represents the scoring matrix of the q -th meta-path prediction. Our optimization goal is to minimize squared error between the predicted score and the real score.

$$\begin{aligned} \min \mathcal{L}(\omega) &= \frac{1}{2} \left\| Y \otimes \left(R - \sum_{q=1}^{|P|} \omega_k^{(q)} \hat{R}^{(q)} \right) \right\|_F^2 + \frac{\lambda}{2} \|\omega\|_F^2 \\ \text{s.t } \omega &\geq 0 \end{aligned} \quad (10)$$

By taking the derivative of the above equation, we can get Eq. (11).

$$\frac{\partial \mathcal{L}(\omega)}{\partial \omega_k^{(q)}} = - \left(Y \otimes \left(R - \sum_{q=1}^{|P|} \omega_k^{(q)} \hat{R}^{(q)} \right) \right) \hat{R}^{(q)T} + \lambda \omega_k^{(q)} \quad (11)$$

where \otimes is *Hadamard product*. Finally, we use the gradient descent algorithm to get the value of $\omega^{(q)}$.

4 Experiment

In this section, we will validate our PTRS model on two real datasets.

4.1 Dataset

In the experiment, we mainly used the movielens movie dataset and the LastFM music dataset. Table 1 details the information about these two data sets.

Table 1. Statistics of MovieLens/LastFM datasets

Dataset	Relation	A	B	A-B
MovieLens	User-Movie	2000	10197	855599
	Movie-Actor	10175	95322	231743
	Movie-Country	10197	10197	10197
	Movie-Director	10156	4061	10156
	Movie-Genre	10197	20	20810
	Movie-Location	10197	7731	49168
	User-Tag	2113	5908	47958
LastFM	User-Artist	1892	17632	92834
	User-User	1892	1892	18802
	Artist-Artist	17632	17632	153399
	User-Tag	1892	11945	184941

4.2 Metrics

We will measure our experimental results using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

$$RMSE = \sqrt{\frac{\sum_{(u,v) \in R_{test}} (R - \hat{R})^2}{|R_{test}|}} \tag{12}$$

$$MAE = \frac{\sum_{(u,v) \in R_{test}} |R - \hat{R}|}{|R_{test}|} \tag{13}$$

where R_{test} denotes the test set. The smaller the RMSE and MAE, the better the experimental results.

4.3 Comparison Method

In order to prove the validity of our proposed PTRS model, we compare the four existing recommendation algorithms for comparative experiments as follows.

UserCF [1] and ItemCF [2]: This is a recommendation algorithm that uses the similarity of users or items.

PMF [16]: This is a recommendation algorithm based on matrix decomposition technique.

HeteRec [10]: This is a recommendation algorithm based on meta-path similarity in HIN.

PTRS: This is our model.

Table 2. Results of effectiveness experiment

Dataset	r	UserCF		ItemCF		PMF		HeteRec		PTRS	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Movielens	0.2	1.5191	1.3412	1.4463	1.2350	1.3153	1.1932	1.2761	1.0933	1.1751	0.9943
	0.4	1.4289	1.2998	1.3532	1.1836	1.2549	1.1525	1.2271	0.9789	1.1049	0.9136
	0.6	1.3852	1.2681	1.3099	1.1426	1.2287	1.1384	1.2027	0.9449	1.0534	0.8565
	0.8	1.3574	1.2312	1.2892	1.1307	1.2164	1.1272	1.1869	0.9308	1.0283	0.8256
LastFM	0.2	1.3130	1.1259	1.2624	1.0736	1.1201	0.9672	1.0359	0.8091	0.9481	0.7351
	0.4	1.2741	1.0961	1.2078	1.0385	1.0572	0.9041	0.9971	0.7863	0.8931	0.6901
	0.6	1.2492	1.0670	1.1800	1.0109	1.0328	0.8921	0.9628	0.7611	0.8627	0.6831
	0.8	1.2296	1.0503	1.1694	0.9996	1.0277	0.8872	0.9531	0.7462	0.8517	0.6798

4.4 Experimental Results

In the experiment, we set different training data ratios ($r = 0.2$ in Table 2 means that the proportion of training data is 20%, and the remaining 80% of the data is used as test data) to verify the validity of our model in different Data sparsity. The test results are analyzed as follows.

- (1) The PMF algorithm is superior to the CF algorithm, which means that the PMF is more accurate in estimating the user and the item by comparing the feature information based on the similarity in the CF.
- (2) HIN-based recommendation algorithm is superior to other algorithms, which also proves that HIN can better analyze user and item feature information. In addition, PTRS algorithm is superior to HeteRec algorithm, which also proves that the embedding of tag information can better analyze the preference information of the user and predict the user's attitude towards the project more accurately. In addition, we validate the effectiveness of the proposed filtering strategy. The experimental results are shown in Fig. 3. In the case of both presence and absence of filtering, our PTRS experimental results have significant differences, which also proves that our strategy for filtering low-quality information in the meta-path is effective.

4.5 Parameter Study

In our proposed model, α represents the weight of recommendation integrated into tag semantics, and the value range of α is $[0, 1]$. Figure 4 illustrates the sensitivity of parameter α on the movielens data set and the LastFM data set. When α is at 0.3, PTRS has the highest accuracy on the movielens dataset; when α is at 0.3–0.4, PTRS has the highest accuracy on the LastFM dataset. This shows that the HIN-based mechanism contributes more in our PTRS model.

Another parameter K is the number of user subgroups clustered by the $k - means$ algorithm. As shown in Fig. 5, the users of the movielens data set preferably cluster 40, while the users of the LastFM data set preferably cluster 60.

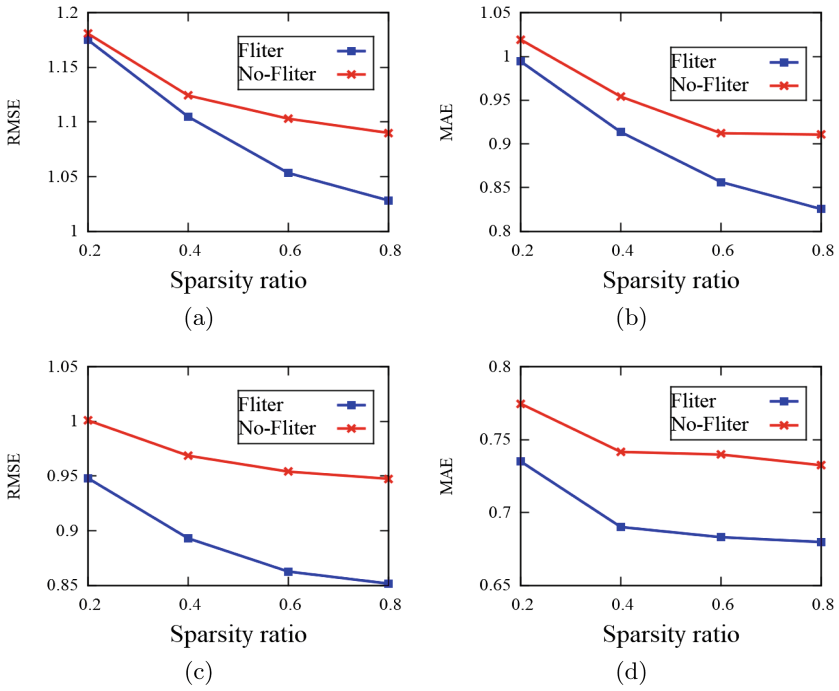


Fig. 3. Comparative experiment of meta-path filtering. (a) movielens, RMSE. (b) movielens, MAE. (c) LastFM, RMSE. (d) LastFM, MAE.

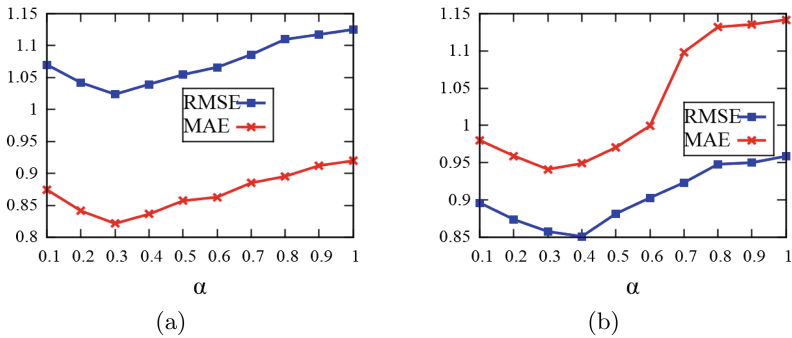


Fig. 4. Performance of PTRS with α . (a) movielens, RMSE, MAE. (b) LastFM, RMSE, MAE.

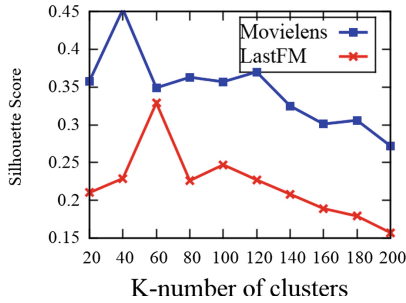


Fig. 5. Cluster number study.

5 Conclusion

In this paper, we integrate the semantic information of tags into the HIN-based recommendation system, and propose a HIN-based recommendation algorithm integrating the semantic information of tags. In this way, users can obtain the semantic relations of items through HIN, and predict users' specific attitudes towards items through the semantic information of tags. In addition, we also design an information filtering strategy, which can filter the influence of low-quality information on recommendation. A large number of experiments prove the effectiveness of PTRS.

Acknowledgement. This work is supported by the National Key R&D Program of China under Grant No. 2017YFB1402102, the National Natural Science Foundation of China under Grant No. 61977044, and the Fundamental Research Funds for the Central Universities of China under Grant Nos. GK201903090, and GK201801004.

References

1. Xu, C., Xu, J., Du, X.: Recommendation algorithm combining the user-based classified regression and the item-based filtering. In: 8th International Conference on Electronic Commerce, pp. 574–578. ICEC, Fredericton (2006)
2. Kim, B., Li, Q., Park, C., Kim, S., Kim, J.: A new approach for combining content-based and collaborative filters. *J. Intell. Inf. Syst.* **27**(1), 79–91 (2006)
3. Zheng, L., Lu, C., Jiang, F., Zhang, J., Philip, S.Y.: Spectral collaborative filtering. In: 12th ACM Conference on Recommender Systems, pp. 311–319. RecSys, Vancouver (2018)
4. Shi, C., Li, Y., Zhang, J., Sun, Y., Philip, S.Y.: A survey of heterogeneous information network analysis. *IEEE Trans. Knowl. Data Eng.* **29**(1), 17–37 (2017)
5. Sun, Y., Han, J., Yan, X., Philip, S.Y., Wu, T.: Pathsim: meta path-based top-k similarity search in heterogeneous information networks. *PVLDB* **4**(11), 992–1003 (2011)
6. Shi, W., Liu, X., Yu, Q.: Correlation-aware multi-label active learning for web service tag recommendation. In: IEEE International Conference on Web Services, pp. 229–236. ICWS, Honolulu (2017)

7. Feng, W., Wang, J.: Incorporating heterogeneous information for personalized tag recommendation in social tagging systems. In: 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1276–1284. KDD, Beijing (2012)
8. Liu, J., Shi, C., Hu, B., Liu, S., Yu, P.S.: Personalized ranking recommendation via integrating multiple feedbacks. In: Kim, J., Shim, K., Cao, L., Lee, J.-G., Lin, X., Moon, Y.-S. (eds.) PAKDD 2017. LNCS (LNAI), vol. 10235, pp. 131–143. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-57529-2_11
9. Shi, C., Zhang, Z., Luo, P., Philip, S.Y., Yue, Y., Wu, B.: Semantic path based personalized recommendation on weighted heterogeneous information networks. In: 24th ACM International Conference on Information and Knowledge Management, pp. 453–462. CIKM, Melbourne (2015)
10. Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B.: Personalized entity recommendation: a heterogeneous information network approach. In 7th ACM International Conference on Web Search and Data Mining, pp. 283–292. WSDM, New York (2014)
11. Zhao, H., Yao, Q., Li, J., Song, Y.: Meta-graph based recommendation fusion over heterogeneous information networks. In: 23th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 635–644. SIGKDD, Halifax (2017)
12. Dong, Y., Chawla, N.V., Swami, A.: metapath2vec: scalable representation learning for heterogeneous networks. In: 23th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 635–644. SIGKDD, Halifax (2017)
13. Fu, T., Lee, W.C., Lei, Z.: HIN2Vec: explore meta-paths in heterogeneous information networks for representation learning. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1797–1806. CIKM, Singapore (2017)
14. Zhu, J., et al.: CHRS: cold start recommendation across multiple heterogeneous information networks. *IEEE Access* **5**, 15283–15299 (2017)
15. Hu, B., Shi, C., Zhao, W., Philip, S.Y.: Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In: 24th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, pp. 1531–1540. KDD, London (2018)
16. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: 21th Annual Conference on Neural Information Processing Systems, pp. 1257–1264. NIPS, Vancouver (2007)
17. Shi, C., Kong, X., Huang, Y., Philip, S.Y., Wu, B.: Hetesim: a general framework for relevance measure in heterogeneous networks. *IEEE Trans. Knowl. Data Eng.* **26**(10), 2479–2492 (2014)



High-Quality Learning Resource Dissemination Based on Opportunistic Networks in Campus Collaborative Learning Context

Peng Li^{1,2,3}, Hong Liu^{1,2,3}, Longjiang Guo^{1,2,3}(✉), Lichen Zhang^{1,2,3},
Xiaoming Wang^{1,2,3}, and Xiaojun Wu^{1,2,3}

¹ Key Laboratory of Modern Teaching Technology,
Ministry of Education, Xi'an 710062, China
longjiangguo@snnu.edu.cn

² Engineering Laboratory of Teaching Information Technology of Shaanxi
Province, Xi'an 710119, China

³ School of Computer Science, Shaanxi Normal University,
Xi'an 710119, China

Abstract. In the campus scenario, a basic community of collaborative teams is formed among the nodes participating in the collaborative learning interaction in the mobile opportunistic network. Due to the existing research does not consider the weak connection, node influence and the contact characteristics between nodes. In this paper, a routing method using a collaborative group as a communication unit is proposed. The route mainly counts the contact characteristics among the groups according to the characteristics of the node movement and predicts the subsequent contact situation. Combined with the weak connection relationship and the node's influence, the optimal route to be transmitted is selected. It has been experimentally verified that the routing method can greatly improve the speed of message dissemination and avoid unnecessary message redundancy and waste of contact opportunities.

Keywords: Opportunistic networks · Collaborative learning · Weak connection · Community influence

1 Introduction

Part of the characteristics of the opportunistic networks comes from the DTN [1] network and MANET, which does not require a complete link connection between the source node and the destination node, but uses a store- carry- forward (SCF) method for message transmission.

Typical routing strategies include Epidemic [2], Delivery [3], Spray and Wait [4], First Contact [5], and ProPHET [6]. Zhao et al. in [7] proposed a community-based route propagation strategy, which establishes a community between nodes and then passes the message to nodes that are active within or between communities, and passes the message to the destination node. In [8], the BSW algorithm only passes the message

to the node with a higher probability of contact than itself. This method will reduce the information redundancy while ensuring the success rate of message delivery. However, there is a possibility that the message may be delayed, and in the above two algorithms, there is no good sense of collaboration within the community, which will lead to a significant reduction in the effectiveness of the community.

In the campus, although the nodes are dense and the movement of the nodes is more regular, there is still a delay in the message propagation in the actual transmission, and since the messages are generally time-sensitive. Thus, this is an urgent problem that how to spread the most valuable learning resources to the nodes of interested as quickly as possible.

This paper focuses on a routing algorithm based on node contact history and node influence [9]. The main contributions are as follows:

- (1) Using the nodes transmission records and the contact relationship among the group community to predict the next contact time, and classify it.
- (2) Prioritize the transmission of communities with high impact, thereby increasing the heterogeneity of the community's message and accelerating the spread of messages.
- (3) When all the nodes to be transmitted have low influence, the best node to be transmitted is selected according to the contact history feature.

It has been experimentally verified that the routing method can greatly improve the speed of message diffusion and avoid unnecessary message redundancy and waste of contact opportunities.

2 Weak Connection in Social Networks

Each learning node on the campus is often in a certain social circle. For example, the familiar learner node will contact the number of times much more than the contact with strange students. In this regard, Granovetter proposed the concept of social relationship intensity [10].

Figure 1 shows a social network relationship with three communities. line indicates the connection relationship between nodes. The dotted line between $a-c$, $d-e$, and $d-f$ indicates the weak connection between them. Although the connection relationship between $a-c$, $d-e$ and $b-f$ is weak, it plays a vital role in the communication between the three communities.

Friedkin proved that weak social relationships contribute to the transmission of academic information between different academic groups [11]. The research in [12, 13] also supports weak social relationships that can help spread information. Literature [14, 15] believes that weak connection is more likely to provide a more diverse variety of re-sources for both parties. Thus, it can be considered that if the node with the weak connection relation-ship is selected, the heterogeneity of the social circle that owns the message is increased and the speed of message diffusion is increased. otherwise, the speed of data diffusion will be delayed to some extent.

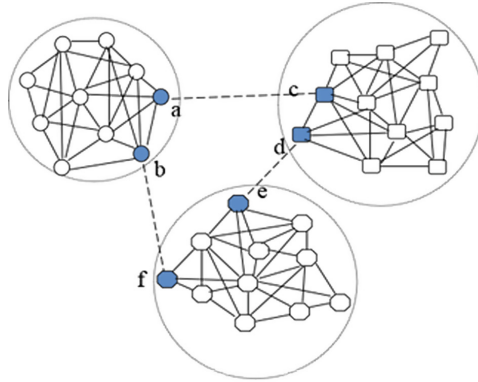


Fig. 1. Community relations in social networks.

The following examples illustrate:

Shown in Fig. 2 are three consecutive time segments, in which the nodes in *a* and *b* groups are in frequent contact, the nodes in *c* and *d* groups are frequently contacted. a_1 is a member of group *a* with learning resources to be transmitted. b_1 , c_1 , and d_1 are members of the *b*, *c*, and *d* groups, respectively, there is no such learning resource. When $t = t_2$, the a_1 node selects to transmit to the b_1 node, and when $t = t_3$, the nodes a_1 , b_1 , c_1 and d_1 connection is interrupted, where b_1 and a_2 meet, and c_1 and d_1 continue to be networked. Since a_2 also has the learning resource, no transmission is performed between b_1 and a_2 , and since c_1 and d_1 do not have the learning resource, therefore no transmission. Therefore the contact opportunity is wasted.

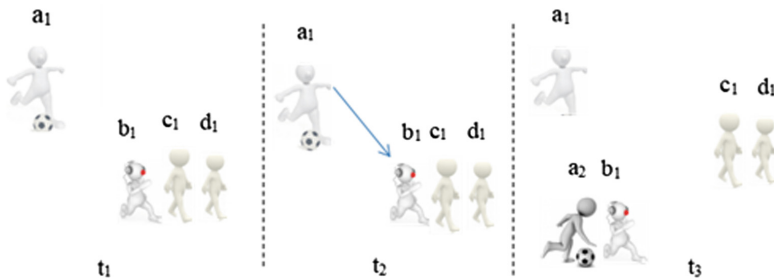


Fig. 2. Closely related nodes as nodes to be transmitted.

The ideal transmission process is shown in Fig. 3 When $t = t_2$, the a_1 chooses to transmit to the c_1 . When $t = t_3$, since a_2 also has the learning resource, a_2 transmits b_1 , and c_1 also has learning resources, so c_1 transmits to d_1 .

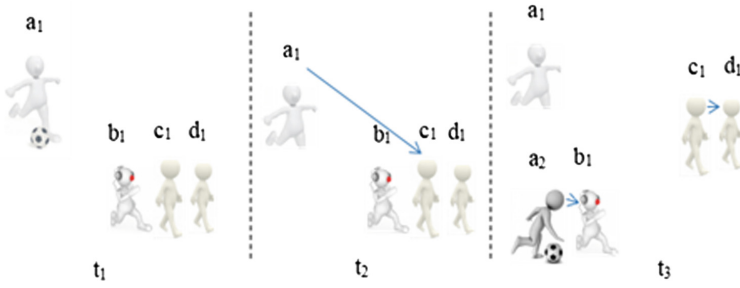


Fig. 3. Weak connection relationship node as the node to be transmitted.

In the campus, the real network contact situation consisting of thousands of collaborative groups is far more complicated than the above case. The above contact situation will also appear in large numbers. In order to optimize the data diffusion process, we propose KSCB algorithm.

3 Contact Interval Duration Estimation and Node Influence

3.1 Estimation and Classification of Contact Interval Duration

The information shown in Table 1 is maintained for each node in the collaborative learning community, and any node in each group community updates its own information table with information held by other nodes. For the next encounter time of any two group nodes, the calculation is mainly based on the contact history between the groups. Since the movement of the nodes in the campus environment shows strong regularity, the predicted value is accurate. The main calculation process is as follows:

Table 1. Node information table.

Symbol	Meaning
Sn	Serial number
id	Contacted group identifier
f	Number of contacts
lmt	The length of time since the last contact
al	Average length of contact interval
emt	Estimated next encounter time
$T_v(v=1, 2, \dots, n)$	Encounter time record
$Tave$	Average contact duration
cud	Estimated value cumulative deviation
Ig	Interest Set
Sn	Serial number

Initialization communication number $f = 0$, last encounter time $T' = 0$.

a. Calculate the number of communications:

$$f = f + 1 \quad (1)$$

b. Calculate the last contact time between groups:

$$lmt = \begin{cases} 0 & f = 1 \\ T - T' & f = 2, 3, \dots, n \end{cases} \quad (2)$$

c. Record encounter time:

$$T' = T \quad (3)$$

d. Calculate the deviation between the estimated time interval and the actual time interval:

$$cud = \begin{cases} 0 & f = 1 \\ cud + lmt & f = 2, 3, \dots, n \end{cases} \quad (4)$$

e. Update the average contact interval:

$$al = \begin{cases} lmt & f = 1 \\ [al \cdot (f - 1) + lmt]/f & f = 2, 3, \dots, n \end{cases} \quad (5)$$

f. Update the average contact duration:

$$T_{ave} = \begin{cases} T - T' & f = 1 \\ [T_{ave} \cdot (f - 1) + T - T']/f & f = 2, 3, \dots, n \end{cases} \quad (6)$$

g. Calculate time interval from the next contact:

$$intv = \begin{cases} lmt & f = 1 \\ \lambda al + (1 - \lambda) \cdot lmt & f = 2, 3, \dots, n \end{cases} \quad (7)$$

h. Estimated time of next contact:

$$emt = \begin{cases} T + intv & f = 1 \\ T + \delta \cdot intv + (1 - \delta) \cdot cud & f = 2, 3, \dots, n \end{cases} \quad (8)$$

According to statistics, when the parameters λ and σ are 0.82 and 0.78 respectively, the estimated value is more accurate.

By counting the contact time intervals of dozens of learner nodes, we can find that there is a certain regularity, as shown in Fig. 4, according to which the values of $intv$ can be divided into six categories: Category I: 0–2 h; Category II: 2–5 h; category III: 5–16 h; Category IV: 16–24 h; Category V: 24–48 h; Category VI: greater than 48 h.

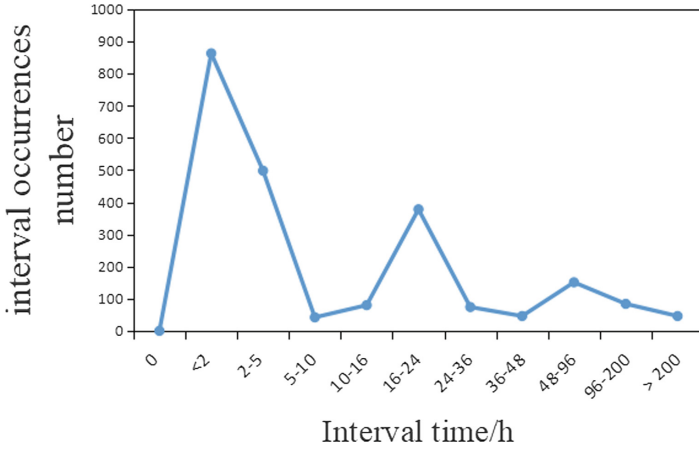


Fig. 4. Contact time interval Statistics.

3.2 Group Influence Model

Because each group has different influences on the external community network, and it can be considered that the influence of any group on other communities is positively related to the team’s ability to contact, and the stronger the contact with other communities, the better the diffusion of information. Therefore, the influence on the group in which the node to be transmitted is located is calculated as follows:

$$T_{AD} = (D_j \cdot D_b)^{S_{AD}} \tag{9}$$

where

$$D_j = \frac{D_g}{G - 1}, S_{AD} = \frac{\sum_{i=1}^{A_g} cud_{Ai}}{A_g \cdot cud_{AD}}, D_b = \sqrt{\frac{1}{D_g} \cdot \sum_{k=1}^{D_g} \left(f_k - \frac{\sum_{i=1}^{D_g} f_i}{D_g} \right)^2}$$

D_j is the community influence ability of collaboration group D , D_b is the contact balance ability of collaboration group D , D_g is the number of groups contacted by group D , and G is the total number of groups. f_i and f_g are the number of connections between the D group and the group with the numbers j and g . T_{AD} is the transmission stability of Group A and Group D , A_g is the number of groups contacted by Group A , and Cud_{Ai} is the cumulative deviation of Group A 's estimate for the group with transmission number i .

When the T_{AD} value is high, it can be considered that the network of the group D to be transmitted has a large influence, and the community that can reach the contact can be effectively transmitted, and the transmission record with the group A is relatively stable.

3.3 Determine the Node to Be Transferred

By counting the connection status of two thousand communication nodes, it can be found that the contact ability of the node is positively correlated with the number of communication times of the node. It can be seen from Fig. 5 that the number of communication times of the node is generally high when the degree of the node is high, and the degree of the node is not necessarily high when the number of node communication is high, so the nodes can be classified into three types. Category I: central nodes, $Max\ sn \geq 150$; Category II nodes, active nodes, $Max\ sn < 150 \ \& \ 600 - 4Max\ sn < \sum_{i=0}^G f_i$. Category III node, common node, $600 - 4Max\ sn \geq \sum_{i=0}^G f_i$. Where G is the total number of groups, f_i is the number of times of communications with group i .

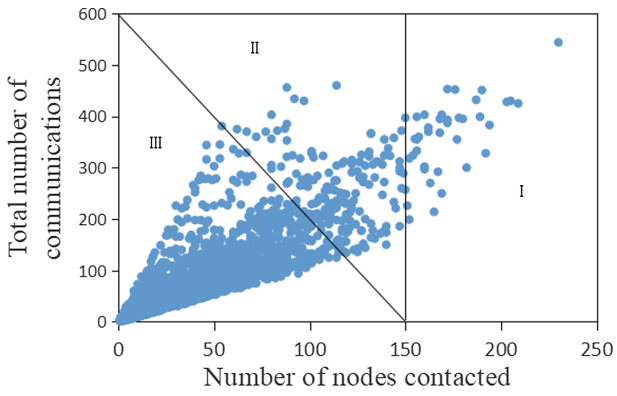


Fig. 5. Number of nodes contacted and number of communication statistics chart.

According to statistics, the proportion of Class I nodes is 1.38%, and the proportion of communication is 9.02%. The proportion of class II nodes is 5.7%, and the proportion of communication is 17.18%. The proportion of Class III nodes is 92.92%, and the proportion of communication is 73.8%. Therefore, it can be considered that the node to be transmitted is a class III node when a weak connection occurs in most cases. Since the class III node does not have obvious network influence capability, when the node to be transmitted is a class III node, the contact history between groups is considered.

The contact history is discussed as follows:

Shown in Fig. 6(a), if the contact opportunities of a and b are similar, and the average contact duration of a is shorter than b , then b is selected as the node to be transmitted.

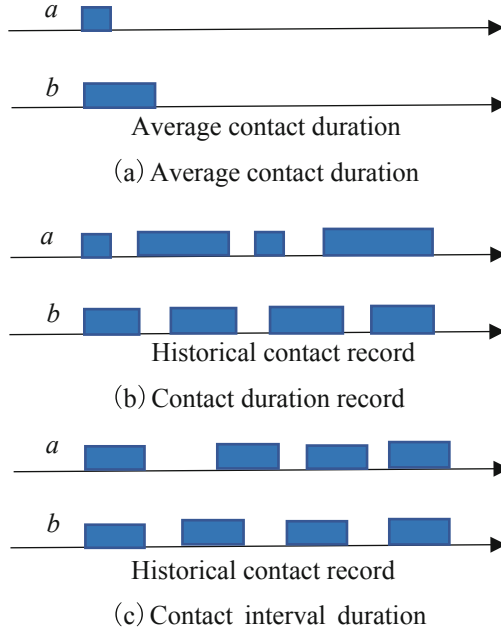


Fig. 6. Contact history comparison char.

Shown in Fig. 6(b), if the average historical contact duration of a is also equivalent to b , but the regularity of the contact duration (T_s) of b is higher, b is selected as the node to be transmitted, calculated as follows:

$$T_s = \sqrt{\frac{1}{f-1} \cdot \sum_{i=1}^f (T_i - T_{ave})^2} \quad (10)$$

As shown in Fig. 6(c), if the degree of contact duration deviation of a is also equivalent to b , but the node contact time interval (al_s) of b is less discrete, b is selected as the node to be transmitted, calculated as follows:

$$al_s = \sqrt{\frac{1}{f-1} \cdot \sum_{i=1}^f (t_i - al)^2} \quad (11)$$

Thus, Algorithm KSCB details as follows:

Algorithm KSCB ai Sequence strategy

```

1:  While  $a_i$  contact with  $Vai(b_i, b_j, c_i, c_j, d_i, \dots, n_i)$  do
2:  Ignore the same group id, get  $Vai_1(b_i, c_i, d_i, \dots, n_i)$ ;
3:  Ignore the nodes that Interest set are incompatible, get  $Vai_2(c_i, \dots, n_i)$ ;
4:  According to  $emt$ , divided the nodes into six categories: I, II, III, IV, V, VI;
5:  The same type of node sorted in descending order according to the  $T$ ;
6:      If ( $Vai_2$ 's  $T$ == III), then
7:          Calculation  $T_{ave}$ , and arranged in ascending order;
8:          If have node with equal  $T_{ave}$ , then
9:              Calculation  $T_s$ , and arranged in ascending order;
10:             If have node with equal  $T_s$ , then
11:                 Calculation  $al_s$ , and arranged in ascending order;
12:                 If have node with equal  $al_s$ ;
13:                     Sort the nodes in descending order according to  $emt$ .
14:                     end if
15:             end if
16:         end if
17:     end if
18: end while

```

4 Simulation Verification and Result Analysis

The infectious disease model is the most widely used model in the field of information dissemination. William and Anderson first proposed the *SIR* model [16], where *S* is healthy but susceptible to infection, Symbol *I* indicates a virus communicator, and *R* is the infected but recovered and antibody-acquired. compared with the above classical model, the node types used in this paper can be divided into four categories, and they are simply referred to as *SEIRS* model, where *S* represents the person to be transmitted (the message is not included, but is interested in this type of message), *E* means indifferent (does not contain this message and not interested), Symbol *I* indicates the message transmitter, *R* means the rejecter, the dynamic model is constructed as follows:

$$\begin{cases} \frac{dS(t)}{dt} = -m \cdot dS(t) \cdot \alpha S(t) - \beta S(t) + \theta R(t) \\ \frac{dE(t)}{dt} = -\gamma E(t) - \delta E(t) + m \cdot dS(t) \cdot \beta I(t) \\ \frac{dI(t)}{dt} = \alpha S(t) + \gamma E(t) - \varepsilon I(t) \\ \frac{dR(t)}{dt} = \delta E(t) + \varepsilon I(t) - \theta R(t) \end{cases} \quad (12)$$

Where $S(t)$, $E(t)$, $I(t)$, and $R(t)$ respectively represent the number of nodes of the to-be-transmitter, the froster, the communicator, and the rejector when time is t . α , β , γ , δ , ε , θ are the corresponding diffusion parameters. Considering the topology of the network in the process of propagation, $m \cdot dS(t)$ is used to indicate the number of related to-be-transmitted nodes, where m represents the ratio of related to-be-transmitted nodes. Assuming that the total number of nodes in the campus is N , there are:

In order to evaluate the performance of the KSCB algorithm, the simulation program is written in the windows environment using the eclipse platform to model the KSCB and epidemic router algorithms. According to the survey conducted by [17], the weak connection relationship accounts for 80% of all interpersonal interactions, and the time spent is 20%, while the strong connection relationship accounts for 20% of all interpersonal relationships. The time spent is 80%. Therefore, it can be assumed that the number of weak connections is τ when the number of intimate contacts is 4τ . Assume the weak connection success probability λ between nodes is 0 to 1 taking random values, and the success probability p of strong connections is 1, and the other simulation parameters are set as Table 2;

Table 2. Simulation environment parameter setting.

Symbol	Meaning	Ranges
T	Simulation time	64 h
N	Total number of campus nodes	10000
P	Proportion of nodes of interest	0.01, 0.02, 0.04, 0.08, 0.2, 0.4, 0.6
t	Message retention time	2 h, 4 h, 6 h, 8 h, 10 h, 20 h

The experimental results are as follows:

It can be seen from Fig. 7 that when the proportion of interested nodes increases, the performance of the KSCB route will increase rapidly. The main reason is that when

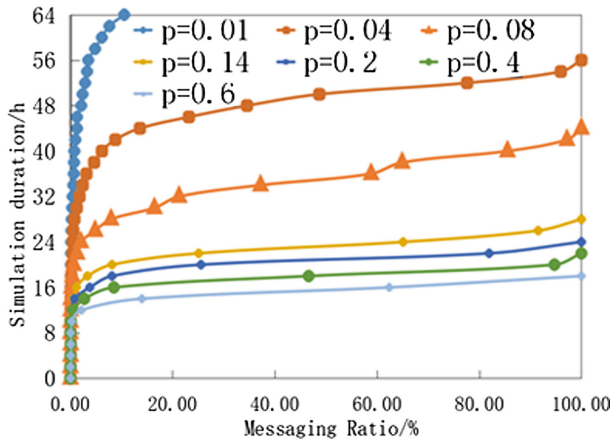


Fig. 7. When $t = 8$ h, the influence of the proportion of interested nodes on message diffusion.

the node of interest is relatively low, the path of information diffusion is less, that is, the link of data transmission between nodes is poor. This leads to lower performance of KSCB routing.

It can be seen from Fig. 8 that when the message storage duration increases, the performance of the KSCB route will increase rapidly, mainly because the node saves the message is short, the number of nodes that have messages and can be delivered is rapidly reduced over time, resulting in reduced performance of KSCB routing.

As can be seen from Figs. 9 and 10, when the proportion of the interest node and the duration of the node save message are moderate, the KSCB route is similar to the

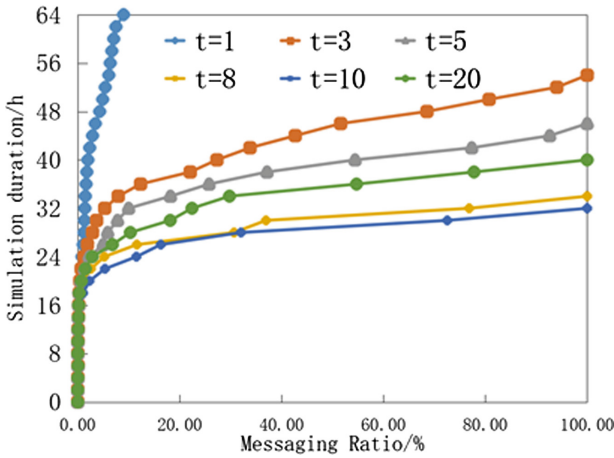


Fig. 8. When $p = 0.08$, the effect of message retention time on message diffusion.

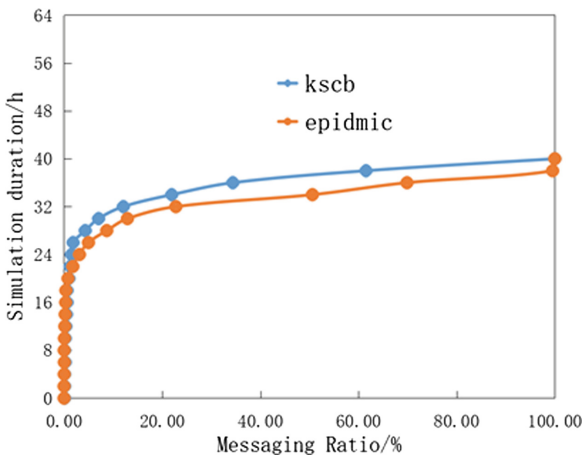


Fig. 9. When $t = 8$ h, $P = 0.08$, the effect of routing mode on the proportions of message delivery.

epidemic route performance. Since KSCB only transmits information to interested nodes, it avoids a large spread of messages, thereby greatly avoiding message redundancy and avoiding waste of contact opportunities for nodes that are not interested in this message.

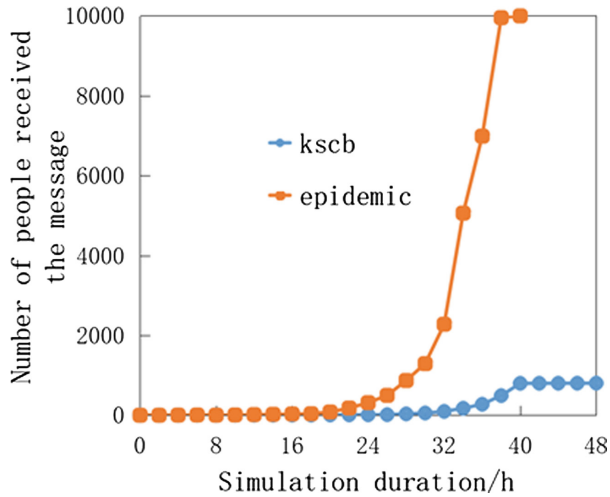


Fig. 10. When $t = 8$ h, $P = 0.08$, the effect of routing on the number of nodes that get messages.

5 Conclusions

In this paper, we predicted the contact situation of node and calculated the influence of the network community of the node. When the community influence of the node to be transmitted is low, the transmission record of the node to be transmitted is considered, and a solution is provided for how to select the most suitable node to be transmitted in the node with similar contact conditions. Simulation experiments show that KSCB routing can achieve better transmission effect when the proportion of nodes of interest and the time of the node saved the information are reasonable, and avoid the redundancy of messages and the waste of transmission opportunities for nodes that are not interested.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 61877037) and the National Natural Science Foundation of China (No. 61977044).

References

1. Fall, K.: A delay-tolerant network architecture for challenged internets. In: Proceedings of the 2003 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, p. 27. ACM, Karlsruhe (2003)

2. Vahdat, A., Becker, D.: Epidemic Routing for Partially-Connected Ad Hoc Networks. *Handbook of Systemic Autoimmune Diseases* (2000)
3. Pelusi, L.: Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. *IEEE Commun. Mag.* **44**(11), 134–141 (2006)
4. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: *ACM SIGCOMM Workshop on Delay-tolerant Networking*, pp. 252–259. ACM, New York (2005)
5. Sushant, J., Kevin, R., Rabin, K.: Routing in a delay tolerant network. In: *Technologies, Architectures, and Protocols for Computer Communication*, pp. 145–158. ACM, Portland (2004)
6. Lindgren, A., Doria, A., Schelén, O.: Probabilistic routing in intermittently connected networks. In: *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 7, no. 3, pp. 19–20 (2003)
7. Zhao, R., Zhang, L., et al.: A novel energy-efficient probabilistic routing method for mobile opportunistic networks. *EURASIP J. Wirel. Commun. Netw.* **2018**(1), 263 (2018)
8. Hui, P., Crowcroft, J., Yoneki, E.: BUBBLE rap: social-based forwarding in delay-tolerant networks. *IEEE Trans. Mob. Comput.* **10**(11), 1576–1589 (2011)
9. Chen, X., Shang, C., Wong, B., et al.: Efficient multicast algorithms in opportunistic mobile social networks using community and social features. *Comput. Netw.* **111**, 71–81 (2016)
10. Petroczi, A., Bacsó, F., et al.: Measuring tie-strength in virtual social networks. *Connections* **91**(1), 39–52 (2006)
11. Friedkin, N.: A test of structural features of Granovetter's strength of weak ties theory. *Soc. Netw.* **2**(4), 411–422 (1980)
12. Rogers, E.M.: Perspectives on social network research. In: *Network Analysis of the Diffusion of Innovations*, pp. 137–164 (1979)
13. Fine, G.A., Kleinman, S.: Rethinking subculture: an interactionist analysis. *Am. J. Sociol.* **85** (1), 1–20 (1979)
14. Kristiansson, S.: Enriching and simplifying communication by social prioritization. In: *International Conference on Advances in Social Networks Analysis and Mining*, Odense, Denmark, pp. 336–340 (2010)
15. Haythornthwaite, W.C., Garton, L.: Studying online social networks. *J. Comput.-Mediated Commun.* **3**(1), 1–5 (1997)
16. May, R.M., Anderson, R.M.: Population biology of infectious disease: part II. *Nature* **280**, 455–461 (1979)
17. Azar, S., Machado, J.C., et al.: Motivations to interact with brands on Facebook - towards a typology of consumer - brand interactions. *J. Brand Manag.* **23**(2), 153–178 (2016)

IntelliSense, Location and Tracking



Integrated Redundant APs Reduction and Transfer Learning for Indoor WLAN Intrusion Detection via Link-Layer Data Transformation

Xinyue Li^(✉), Mu Zhou, Yaoping Li, Hui Yuan, and Zengshan Tian

School of Communication and Information Engineering,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, China

443126330@qq.com, {zhoumu, tianzs}@cqupt.edu.cn,
1271194359@qq.com, 1362168887@qq.com

Abstract. Indoor intrusion detection technology has been widely used in smart home management, public safety, disaster relief, and other fields. In recent years, with the rapid deployment of Wireless Local Area Network (WLAN) and general support of the IEEE 802.11 protocol by mobile devices, indoor intrusion detection can be realized conveniently. Most of the existing indoor intrusion detection algorithms have large computational and storage overheads and do not consider the instability of signals in the indoor environment. In response to this compelling problem, this paper proposes a new integrated redundant Access Points (APs) reduction and transfer learning for indoor WLAN intrusion detection via link-layer data transformation. First, the detection technology for mobile APs based on a fuzzy rough set is exploited to filter the redundant APs in the indoor environment. Second, the target domain and the source domain are constructed through the link-layer data of the online phase and the offline phase. Then, the Maximum Mean Deviation (MMD) minimum value corresponding to the two domains is worked out by the mathematical statistics method to obtain the optimized migration matrix, and the link-layer information of the two domains is transferred into the same subspace by using the matrix. Finally, the optimal intrusion detection classifiers are obtained by training the transferred link-layer data. This method not only has better robustness in the complex indoor environment but also reduces time and labor costs.

Keywords: Intrusion detection · Redundant APs reduction · Transfer learning · Link-layer data · Subspace

1 Introduction

With the rapid development of wireless networks, it is more convenient to experience various services. Global Positioning System (GPS), sensors, some kinds of rays, computer-vision, and Wireless Local Area Network (WLAN) are widely exploited in existing indoor target intrusion sensing techniques [1]. Among them, video image-based techniques tend to expose some critical information about the target [2], which

also has specific requirements for lighting conditions meanwhile. Techniques based on GPS and various rays often require hardware facilities [3], and it is inevitable for targets to carry special equipment, which may increase the complexity of such methods. Although sensor-based techniques do not require the deployment of monitoring equipment or the purchase of expensive high-precision devices [4], their performance tends to depend on the deployment of sensors in the test environment, which will lead to higher labor costs. However, techniques based on WLAN pays more attention to the protection of sensitive features of the target and have higher robustness. Besides, WLAN devices are more comfortable to deploy and less expensive, which are nearly unrestricted by Line-of-sight (LOS) [5]. Recently, thanks to the fine-grained description of channel performance, Channel State Information (CSI) has attracted full attention from many researchers and has been applied to improve indoor localization accuracy. However, the CSI-based intrusion detection system [6] usually has a high complexity which will make it challenging to work out. Therefore, in order to improve the performance of the intrusion detection system, WLAN-based intrusion detection techniques are increasingly applied in most aspects of people's lives, such as company database maintenance, airport tower monitoring, and campus network management.

In this paper, we exploit the joint decision criterion based on fuzzy rough set [7] to filter out redundant APs among mobile APs, and then propose to construct source and target domains by exploiting the link-layer information labeled in the online phase and the link-layer information unlabeled in the offline phase respectively. After that, we construct the transfer matrix by calculating Maximum Mean Discrepancy (MMD) of their edge distributions [8], which can transfer the information in these two domains into the same subspace. Finally, we can obtain more stable intrusion detection by training the data in this subspace. The rest of this paper is structured as follows. The second section introduces some research on indoor WLAN intrusion detection algorithms. Section 3 details the method proposed in this paper, and then the associated experimental results are given in the fourth section. Finally, the fifth part summarizes the paper and discusses future research directions.

2 Related Work

In an open wireless network communication environment, researchers have conducted extensive research on indoor signals through collecting, measurement, and modeling. However, due to the openness and complexity of wireless transmission networks, it is vulnerable for systems to be invaded by harmful data. Researchers have to spend vast amounts of time and labor on reducing intrusion interference when studying indoor wireless signals [8]. With the tremendous advancement in computer processing, data processing methods based on machine learning (such as transfer learning) have recently been widely used in intrusion detection systems. Among them, WLAN-based passive intrusion detection is a novel type of technology, which is suitable for many applications such as intrusion detection, smart home, and border protection [9]. AR-Alarm

[10] has designed and implemented an adaptive and robust intrusion detection system by using fine-grained CSI in commercial WLAN devices. In order to reduce the complexity of intrusion detection, the authors in [11] propose an implementation of an intrusion detection system for WLAN networks using an ensemble learning method, in which the AWID WLAN intrusion dataset is used to discover the necessary features needed for the efficient Intrusion Detection System (IDS) implementation. In order to solve the poor privacy and deployment of specialized devices of intrusion detection systems, the authors in [12] propose WLID, a whole-home level intrusion detection system based on Received Signal Strength (RSS) measurements of WLAN in the indoor complex environment. WLID can be connected to WLAN-enabled devices such as network TVs and smart home appliances, which realizes high-precision intrusion detection through the detection algorithm of non-parametric statistical methods. Different from the above methods, this paper proposes a new integrated redundant Access Points (APs) reduction and transfer learning for indoor WLAN intrusion detection via link-layer data transformation, which reduces the influence of signal time variability in the indoor environment by constructing the transfer matrix of the source domain and target domain. In a word, the two main contributions of this paper are as follows.

- The impact of signal fluctuation is minimized by calculating the MMD of the marginal distributions that is made up of the labeled link layer data and the unlabeled link layer data.
- The minimum value of MMD is used to construct the optimal transfer matrix, which can transfer the link-layer data to the same subspace, and then it is more convenient to obtain the classifiers that make the intrusion detection system more accurate and stable by training these data.

3 System Description

The overview of the proposed system structure is shown in Fig. 1. First, several WLAN APs and MPs are deployed in the test area. Second, the redundant APs are removed according to the joint judgment criterion based on fuzzy rough among the mobile APs. Third, the source domain and target domain are constructed by labeled RSS data in the offline phase and unlabeled RSS data in the online phase respectively, and then the minimum values of the MMD of their marginal distributions are calculated. Finally, the optimized migration matrix based on the MMD minimum is obtained, and then the RSS data are transferred into the same subspace to achieve the training of intrusion detection classifiers.

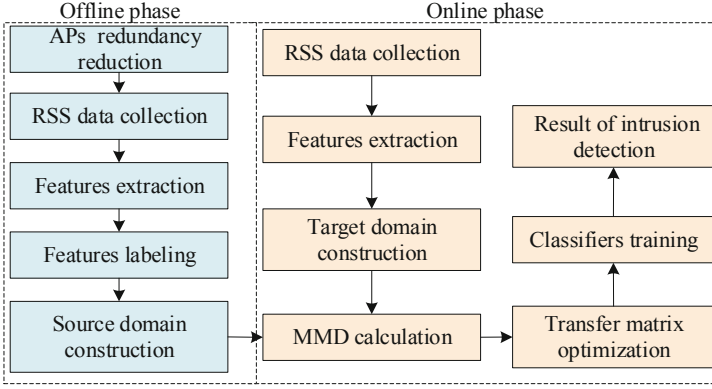


Fig. 1. System structure.

3.1 Redundant APs Reduction

In the indoor environment, too many mobile APs will not improve the performance of the intrusion detection system, which usually increases the computing and storage cost. Therefore, it is necessary to filter out the redundant APs in the test environment before intrusion detection, which will significantly improve the effectiveness of the remaining APs. The fingerprint database is denoted as \mathbf{F} , where the elements in each row stands for the received signal strength at a certain RP. The data of received signal strength from the h APs which do not move is stored in the first h columns, and the 2-D dimension position of RPs is stored in the last two columns, as shown in formula (1).

$$\mathbf{F} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1h} & a_1 & b_1 \\ x_{21} & x_{22} & \cdots & x_{2h} & a_2 & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nh} & a_n & b_n \end{bmatrix} \quad (1)$$

where \mathbf{F} is represented as a fuzzy rough set, in which condition and decision attributes are the RSS data of the h stationary APs and corresponding RPs coordinates. Out of universality, since RPs coordinates have an infinite number of possible values, RPs is also represented by an infinite number of labels. To reduce the complexity of the problem, we divide the experimental scenario into c subregions, and each RP is marked as the ID of the subregion to which it belonged. Then, \mathbf{F} is transferred into

$$\mathbf{F}' = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1h} & a_1 & d_1 \\ x_{21} & x_{22} & \cdots & x_{2h} & a_2 & d_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nh} & a_n & d_n \end{bmatrix} \quad (2)$$

and the equivalent target matrix is written as

$$M(R) = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} \tag{3}$$

where $r_{ij}(\in [0, 1])$ is the correlation value that indicates the relations between the i -th RP and the j -th RP. And R is denoted as the fuzzy equivalence relations that indicate the reflectance, symmetry, and transitive relations.

The concept of fuzzy information entropy is often used in wireless network detection. In this paper, we use a fuzzy rough set to exclude inferior APs. The cardinality of $[x_i]_R$ is defined as

$$|[x_i]_R| = \sum_{j=1}^c r_{ij} \tag{4}$$

and then the information quantity of fuzzy equivalence relation is calculated as

$$H(R) = -\frac{1}{n^2} \sum_{j=1}^n |[x_i]_R| \tag{5}$$

We divide all AP collections (or sets of conditional attributes) A into two subsets B_1 and B_2 and the corresponding fuzzy equivalence classes are represented as $[x_i]_{B_1}$ and $[x_i]_{B_2}$. Then, their joint entropy are calculated as

$$H(B_1, B_2) = -\frac{1}{n^2} \sum_{j=1}^n \log_2 \frac{|[x_i]_{B_1} \cap [x_i]_{B_2}|}{n} \tag{6}$$

Second, the impurity of B_2 under the condition of B_1 is defined as

$$H(B_2|B_1) = -\frac{1}{n^2} \sum_{j=1}^n \log_2 \frac{|[x_i]_{B_2} \cap [x_i]_{B_1}|}{|[x_i]_{B_1}|} \tag{7}$$

when $a \in A$. For all of the access points belonging to the subset of A , the importance degree of a with the subdomain labels d can be worked out by

$$\text{Sig}(a, B, d) = H(d|B - a) - H(d|B) \tag{8}$$

Finally, combining with appropriate optimization algorithms, the redundant APs can be identified from the raw set of APs in the indoor environment.

3.2 Source and Target Domains Construction

During the offline stage, the set of received signal strength vectors gathered at n MPs from m APs is denoted as $\mathbf{R}_S = \{\mathbf{r}_1^1, \dots, \mathbf{r}_m^1, \dots, \mathbf{r}_1^n, \dots, \mathbf{r}_m^n\}$, and $\mathbf{r}_j^i (i = 1, \dots, n; j = 1, \dots, m)$ represents the received signal strength gathered at the i -th MP from the j -th AP. Then, the offline sliding window size is represented as L , which can be utilized to extract some important information of the received signal strength vectors including the mean of them, the probability of the received signal strength exceeding the mean, variance, extreme value, propagation scope, median and received signal strength with the highest probability of occurrence. Based on these data, the feature matrix is set up as $\mathbf{X}_S = (\mathbf{x}_1, \dots, \mathbf{x}_{n_s})^T \in \Omega^{n_s \times p}$, where $\mathbf{x}_s = (x_1^s, \dots, x_p^s)$ ($s = 1, \dots, n_s$) represents the features of the received signal strength vectors in the s -th offline sliding window, n_s is the number of offline sliding windows, and $p (= 8 \text{ nm})$ is the number of RSS characteristics that is equal to the number of types of feature which determined how many pairs of MPs and APs we deployed. Besides, the number of circumstance states can be denoted as K (including 1 silence and $K - 1$ areas intrusion states), and the label of the s -th sliding window corresponding to the k -th ($k = 1, \dots, K$) environment state is represented as $y_s = k$ where the set of special marks of sliding windows during the offline stage can be obtained as $\mathbf{y}_S = (y_1, \dots, y_{n_s})^T$, and then the source domain can be formed as $\mathcal{D}_S = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_{n_s}, y_{n_s})\}$. Similarly, the source domain can be formed as $\mathcal{D}_T = \mathbf{X}_T = \{\mathbf{x}'_1, \dots, \mathbf{x}'_{n_t}\}^T$.

3.3 Transfer Matrix Construction

In order to analyze the problem conveniently, it is reasonable to suppose that the RSS characteristics in online and offline phase follow the same distribution, that is to say, $P(\mathbf{X}_S) = P(\mathbf{X}_T)$, which is also corresponding to the same conditional distribution $Q(\mathbf{y}_S | \mathbf{X}_S) = Q(\mathbf{y}_T | \mathbf{X}_T)$, where \mathbf{y}_T represents the set of special marks of sliding windows during the online stage. Therefore, the empirical measuring of the difference between these two marginal distributions can be worked out by

$$D(P(\mathbf{X}_S), P(\mathbf{X}_T)) = \left\| \frac{1}{n_s} \sum_{s=1}^{n_s} \phi(\mathbf{x}_s) - \frac{1}{n_t} \sum_{t=1}^{n_t} \phi(\mathbf{x}'_t) \right\|_{\mathcal{H}}^2 \quad (9)$$

where the notation “ $\|\cdot\|_{\mathcal{H}}^2$ ” represents the two norm calculation in Reproducing Kernel Hilbert Space (RKHS) and $\phi(\cdot)$ represents the transfer function which can transfer the RSS characteristics into the data in RKHS. In fact, the difference above is also named MMD of marginal distributions of the RSS data in source and target domains, which play an important role in indicating the inherent variety between the data in source domain and target domain. By rewriting (9) into matrix form (we can see (10)), the problem of working out $\phi(\cdot)$ is the same to the one of optimizing transfer matrix \mathbf{W} .

$$D(P(\mathbf{X}_S), P(\mathbf{X}_T)) = \text{tr}(\mathbf{K}\mathbf{W}\mathbf{W}^T\mathbf{K}\mathbf{L}) = \text{tr}(\mathbf{W}^T\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W}) \quad (10)$$

where the notation tr represents the operation of tailing after targets, $\mathbf{K} = \begin{bmatrix} \mathbf{K}_{ss} & \mathbf{K}_{st} \\ \mathbf{K}_{ts} & \mathbf{K}_{tt} \end{bmatrix} \in \Omega^{(n_s+n_t) \times (n_s+n_t)}$ is the kernel matrix where the elements in the i -th row and j -th column of \mathbf{K}_{ss} , \mathbf{K}_{st} , and \mathbf{K}_{tt} are $(\mathbf{K}_{ss})_{ij} = f(\mathbf{x}_i, \mathbf{x}_j)$ in which $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_S$, $(\mathbf{K}_{st})_{ij} = f(\mathbf{x}_i, \mathbf{x}'_j)$ in which $\mathbf{x}_i \in \mathbf{X}_S, \mathbf{x}'_j \in \mathbf{X}_T$, and $(\mathbf{K}_{tt})_{ij} = f(\mathbf{x}'_i, \mathbf{x}'_j)$ in which $\mathbf{x}'_i, \mathbf{x}'_j \in \mathbf{X}_T$ respectively and $\mathbf{K}_{st} = \mathbf{K}_{ts}^T$, $f(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / \gamma\right)$ is the Gaussian kernel function with the bandwidth equaling to the average of the pairwise distances between offline and online received signal strength data [13], $q (< p)$ is the space size of the information conveyed from the characteristics of received signal strength, and the component in the i -th row and j -th column of parameter matrix \mathbf{L} is

$$(\mathbf{L})_{ij} = \begin{cases} 1/n_s^2 & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_S \\ 1/n_t^2 & \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_T \\ -1/n_s n_t & \text{Otherwise} \end{cases}$$

Then, to improve the transfer matrix \mathbf{W} , we can consider the topic of MMD minimization as

$$\begin{aligned} \min_{\mathbf{W}} \quad & \text{tr}(\mathbf{W}^T\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W}) + \lambda \text{tr}(\mathbf{W}^T\mathbf{W}) \\ \text{s.t.} \quad & \mathbf{W}^T\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W} = \mathbf{I} \end{aligned} \quad (11)$$

where λ is the tradeoff coefficient, \mathbf{I} is the unit matrix, $\mathbf{H} = \mathbf{I} - \mathbf{e}\mathbf{e}^T / (n_s + n_t)$, \mathbf{e} is the all-one column vector, and $\text{tr}(\mathbf{W}^T\mathbf{W})$ is the regular term.

To work out the problem of MMD minimization in (11), the Lagrangian method is usually applied to form

$$L(\mathbf{W}) = \mathbf{W}^T\mathbf{K}\mathbf{L}\mathbf{K}\mathbf{W} + \lambda \mathbf{W}^T\mathbf{W} + (\mathbf{I} - \mathbf{W}^T\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W})\boldsymbol{\varphi} \quad (12)$$

where $\boldsymbol{\varphi}$ represents the diagonal matrix which is made up of Lagrangian multipliers. We work out the partial derivative of $L(\mathbf{W})$ in regard to \mathbf{W} to obtain

$$\partial L(\mathbf{W}) / \partial \mathbf{W} = 2(\mathbf{K}\mathbf{L}\mathbf{K} + \lambda \mathbf{I})\mathbf{W} - 2\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W}\boldsymbol{\varphi} \quad (13)$$

By setting $\partial L(\mathbf{W}) / \partial \mathbf{W} = 0$, it is easy to obtain

$$(\mathbf{K}\mathbf{L}\mathbf{K} + \lambda \mathbf{I})\mathbf{W} = \mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W}\boldsymbol{\varphi} \quad (14)$$

Further, we multiply both sides of (14) with \mathbf{W}^T to obtain

$$\mathbf{W}^T(\mathbf{K}\mathbf{L}\mathbf{K} + \lambda\mathbf{I})\mathbf{W} = \mathbf{W}^T\mathbf{K}\mathbf{H}\mathbf{K}\mathbf{W}\boldsymbol{\phi} = \boldsymbol{\phi}\mathbf{I} \tag{15}$$

From the equations above, we can see that the solution to the problem of MMD minimization (or called optimal transfer matrix \mathbf{W}) can be worked out from the q generalized vectors of characteristics corresponding to the q non-zero minimum generalized vectors of characteristics of $(\mathbf{K}\mathbf{L}\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\mathbf{H}\mathbf{K}$. Finally, we utilize the optimal transfer matrix \mathbf{W} to transfer the link-layer data in both source and target domains into the data in the same subspace, and then the classifiers exploited in intrusion detection are trained through these data to reduce the instability of the proposed method.

4 Experiment Results

4.1 Simulation Environment

The layout of simulation environment is shown in Fig. 2, where we represent the MP as MP1 and the four APs as AP1, AP2, AP3, and AP4. The logarithmic attenuation transmission model is applied in this simulation, and we set the noise to follow the Gaussian distribution. Under both the silence and intrusion states the test area is divided into four areas that can be denoted as a1, a2, a3, and a4. In the following results, we select the probability of judging as intrusion state in the absence of intrusion (or called False Positive (FP)), probability of judging as silence state in the presence of intrusion (or called False Negative (FN)), and probability of correct states judgment (or called Detection Accuracy (DA)) as three indices to study the performance of intrusion detection approach proposed in this paper.

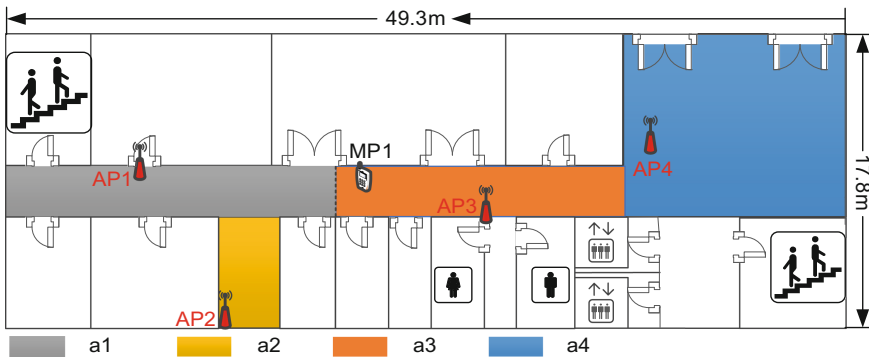


Fig. 2. Layout of simulation environment.

4.2 Simulation Results

To perform the experiments conveniently, We assume that the communication range of the APs deployed can cover the test area. Figure 3 illustrates the confusion matrix

about five circumstance states that include one silence and four areas intrusion states, where the element in the i -th row and j -th column means the probability of judging the i -th circumstance state as the j -th one. It can be seen that the more significant the diagonal element in the confusion matrix is, the better the performance of the intrusion detection system will be. As the color changes from blue to red, it indicates that the accuracy of intrusion detection is gradually increasing. The size of the sliding window also influences the performance of intrusion detection. Generally speaking, appropriately increasing the size of the sliding window can make the intrusion detection classifiers acquire more compelling features of the RSS data, thereby enhancing system robustness. However, excessive size of the sliding window will increase the judgment delay of the system.

Besides, in Table 1, the FP, FN, and DA of the proposed approach are compared with the ones of other three well-performed intrusion detection approaches (i.e., RASID [14], PNN [15], and PRNN [16]). It can be seen that the proposed method has lower FP and FN on the one hand and higher DA on the other hand, which means that the proposed method optimizes the performance of the intrusion detection system. In fact, PNN and PRNN are not sensitive to time-varying noise in the environment, which may result in their inaccurate analysis of RSS data. In RASID, the silence state in the offline phase may be affected by environmental fluctuation.

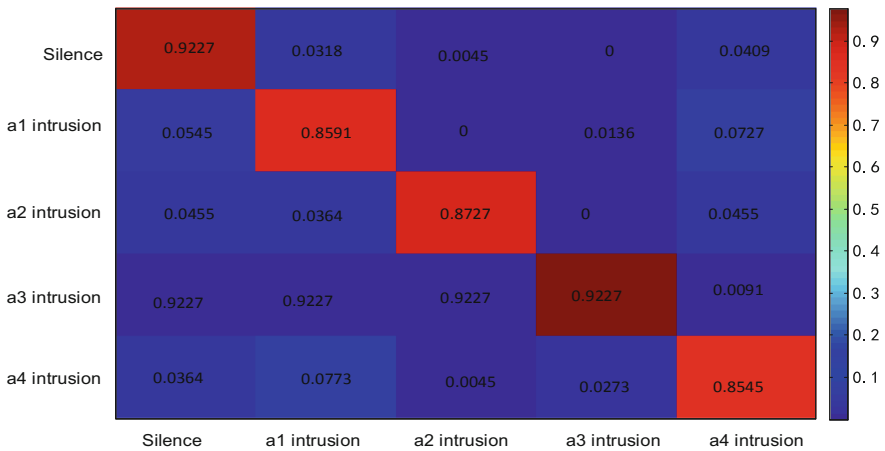


Fig. 3. Confusion matrix in simulation environment.

Table 1. FP, FN, and DA by different intrusion detection approaches.

Metrics	RASID	PNN	PRNN	Proposed
FP	6.87%	3.49%	0	4.21%
FN	3.28%	2.31%	0	0
DA	93.61%	94.39%	95.51%	97.62%

4.3 Actual Environment

The actual environment is shown in Fig. 4, in which there are three MPs (SAMSUNG GT-S7568) which can be denoted as MP2, MP3, and MP4 and five APs (D-Link DAP 2310) which can be denoted as AP5, AP6, AP7, AP8, and AP9. Different from the layout of simulation environment, the experiments are conducted in an actual indoor WLAN environment which is made up of a corridor and a corner. This environment is divided into four areas (notated as a5, a6, a7, and a8), in each of which the RSS data are collected under both the silence and intrusion states with the sampling rate 2 Hz. 600 sets of RSS data corresponding to all pairs of AP and MP are stored in our database. In the following results, we also select the probability of FP, FN, and DA as three indices to study the performance of the proposed approach.

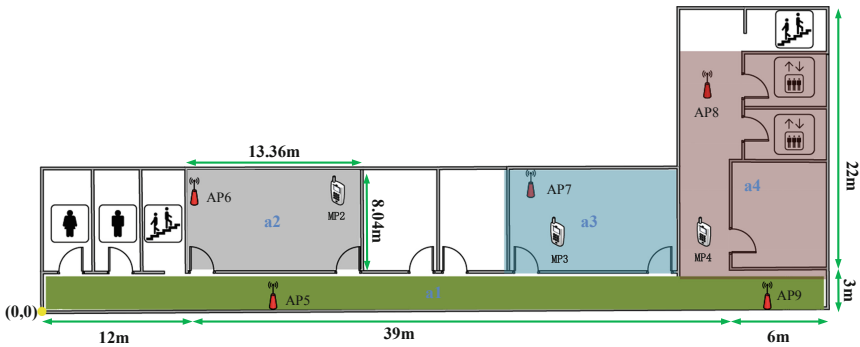


Fig. 4. Layout of actual environment.

4.4 Actual Results

Figure 5 illustrates the confusion matrix with different sliding window sizes, which also includes five circumstance states that include one silence and four areas intrusion states. The size of sliding window is denoted as L . It can be seen that the size of the sliding window not only affects the value of the diagonal elements but also affects the judgment of the silence state. The dimensions of the optimized transfer matrix can be extended by increasing the size of the sliding window, which can make us obtain more feature vectors.

Besides, in Table 2, we conduct experiments in ASID, PNN, PRNN, and the system proposed. We can see clearly that the FP of the proposed method is 4.08%, which is lower than that of RASID. On the other hand, the DA of the proposed method is 97.93%, which is the highest of the four methods. Based on the performance, the method in this paper not only improves the performance of the intrusion detection system but also enhances the robustness of the system.

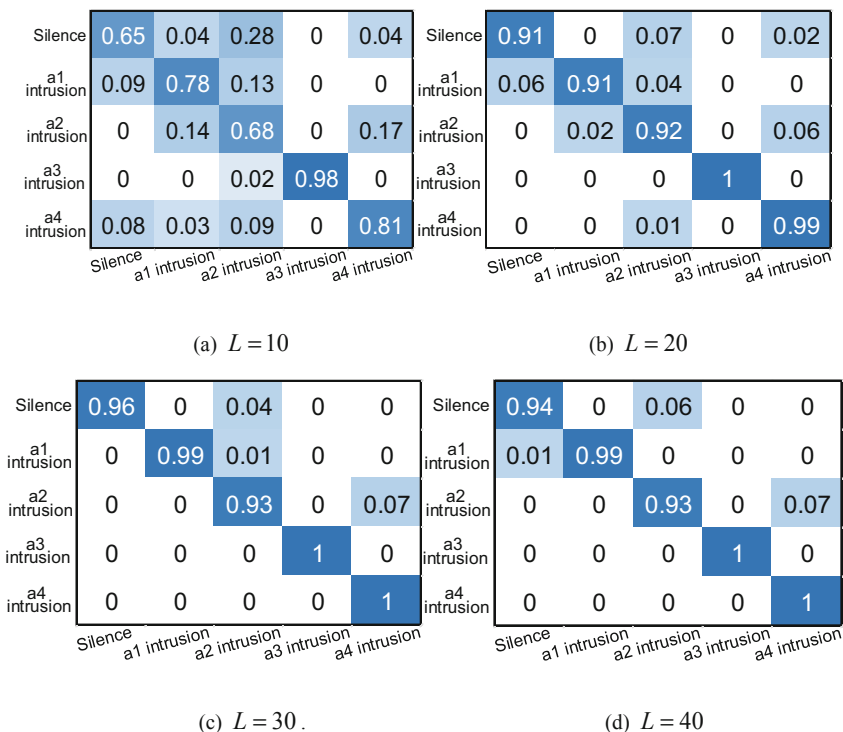


Fig. 5. Confusion matrices with different sliding window size.

Table 2. FP, FN, and DA by different intrusion detection approaches.

Metrics	RASID	PNN	PRNN	Proposed
FP	5.49%	3.34%	0	4.08%
FN	3.62%	2.47%	0	0
DA	93.79%	94.21%	95.48%	97.93%

5 Conclusion

This paper proposes a new integrated redundant APs reduction and transfer learning method for indoor WLAN intrusion detection. Different from the traditional intrusion detection system, we first preprocess the mobile APs in the test environment, and then exploit the joint decision criterion based on fuzzy rough set to filter out redundant APs. Second, the optimal migration matrix is obtained by solving the minimum MMD of the edge distribution of the source and target domains. Third, we utilize this matrix to transfer the RSS data set of the source and target domains into the same subspace, and finally train the RSS data in this subspace to obtain the optimized classifier of the intrusion detection system.

In summary, the proposed method reduces the computation and deployment overhead of the intrusion detection system to a certain extent, and also effectively

improves the stability of the system. Further research work can focus on solving the problem of multi-target intrusion when the indoor signal propagation characteristics are more complex. Besides, when the multi-target motions interfere with each other, using the statistical characteristics between the target positions to improve the localization accuracy of the target is also a problem worthy of attention.

Acknowledgement. This work is supported in part by the National Natural Science Foundation of China (61771083, 61704015), Program for Changjiang Scholars and Innovative Research Team in University (IRT1299), and Fundamental Science and Frontier Technology Research Project of Chongqing (cstc2017jcyjAX0380).

References

1. Moreau, J., Ambellouis, S., Ruichek, Y.: Fisheye-based method for GPS localization improvement in unknown semi-obstructed areas. *Sensors* **17**(1), 119 (2017)
2. Nam, J.Y., Rao, K.R.: Image coding using a classified DCT/VQ based on two-channel conjugate vector quantization. *IEEE Trans. Circuits Syst. Video Technol.* **1**(4), 325–336 (2015)
3. Yi, L., Zhao, H., Yue, H., et al.: Integration of vision and topological self-localization for intelligent vehicles. *Mechatronics* **51**, 46–58 (2018)
4. Cichon, D., Psiuk, R., Brauer, H., et al.: A hall-sensor-based localization method with six degrees of freedom using unscented Kalman filter. *IEEE Sens. J.* **99**, 1 (2019)
5. Fan, J., Awan, A.S.: Non-line-of-sight identification based on unsupervised machine learning in ultra wideband systems. *IEEE Access* **7**(99), 1 (2019)
6. Musavi, S.A., Hashemi, M.R.: HPCgnature: a hardware-based application-level intrusion detection system. *IET Inf. Secur.* **13**(1), 19–26 (2019)
7. Beaubouef, T., Petry, F.E.: Fuzzy rough set techniques for uncertainty processing in a relational database. *Int. J. Intell. Syst.* **15**(5), 389–424 (2000)
8. Tian, W., Qun, L., Bucci, D.J., et al.: K-medoids clustering of data sequences with composite distributions. *IEEE Trans. Signal Process.* **67**(8), 2093–2106 (2019)
9. Yi, Z.: Optimized detection algorithm of complex intrusion interference signal in mobile wireless network. *J. Discrete Math. Sci. Crypt.* **21**(3), 771–779 (2018)
10. Dey, S., Hossain, A.: Session-key establishment and authentication in a smart home network using public key cryptography. *IEEE Sensors Lett.* **3**(4), 1–4 (2019)
11. Qi, W., Yi, Z., Jing, G.: Target positioning algorithm based on WSN in perimeter intrusion detection. *Comput. Eng.* **39**(9), 39–44 (2013)
12. Singh, R., Kumar, H., Singla, R.K., et al.: Internet attacks and intrusion detection system. *Online Inf. Rev.* **41**(2), 171–184 (2017)
13. Shang, Z., Da, C., Qiao, X., et al.: Optimizing the Gaussian kernel function with the formulated kernel target alignment criterion for two-class pattern classification. *Pattern Recogn.* **46**(7), 2045–2054 (2013)
14. Kosba, A.E., Saeed, A., Youssef, M.: RASID: a robust WLAN device-free passive motion detection system. In: *IEEE International Conference on Pervasive Computing and Communications*, pp. 180–189, March 2012
15. Fattah, M.A., Ren, F.: GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Comput. Speech Lang.* **23**(1), 126–144 (2009)
16. Mandic, D.P., Chambers, J.A.: Toward an optimal PRNN-based nonlinear predictor. *IEEE Trans. Neural Netw.* **10**(6), 1435–1442 (1999)



A Near-Optimal Heterogeneous Task Allocation Scheme for Mobile Crowdsensing

Guangsheng Feng¹, Quanming Li¹, Junyu Lin^{2(✉)}, Hongwu Lv¹,
Huiqiang Wang¹, and Silin Lv¹

¹ College of Computer Science and Technology, Harbin Engineering University,
Harbin 150001, China

{fengguangsheng, liquanming, lvhongwu, wanghuiqiang, lvsilin}@hrbeu.edu.cn

² Institute of Information Engineering, Chinese Academy of Sciences,
Beijing 100093, China
linjunyu.cas@gmail.com

Abstract. We study the problem of heterogeneous task assignment in mobile crowdsensing (MCS) scenarios where the opportunistic mode and participatory mode coexist. Workers in opportunistic mode complete tasks during their daily routines while workers in participatory mode complete tasks by moving to designated locations. This problem can be simplified into a Knapsack problem which is NP-hard. Then, to solve this problem, we propose a two-phase task assignment algorithm MSHTA based on the workers' mobility and historical information which leverage the advantages of two sensing modes in sensing quality and sensing cost of tasks. Specifically, a task is optimally assigned to workers who meet their sensing requirements (e.g., sensing time, sensing sensor) at each phase. Extensive simulation results show the effectiveness of our proposed algorithm in terms of tasks' sensing quality and tasks' sensing cost.

Keywords: Mobile crowdsensing · Heterogeneous tasks · Hybrid sensing mode

1 Introduction

The popularity of 4G/5G networks and the explosive growth of mobile devices with numerous sensors have enabled a novel sensing paradigm, namely mobile crowdsourcing (MCS) [1]. MCS uses mobile devices held by pedestrians (called workers) as basic sensing units to monitor environments (e.g., traffic congestion [2], air quality [3], noise level [4], etc.) of urban area in real-time. In MCS, workers collect data according to the requirements of tasks assigned to them by MCS platform, and then return the data to MCS platform. By analyzing the data submitted by workers, MCS platform can obtain environmental information of urban areas. Due to the differences of workers, they may submit different data

for the same task where the data could significantly impact the analysis results of MCS platform. Hence, task assignment becomes a major issue in current MCS research works.

Recently, there have been many studies on MCS task assignment [5–12], which can be divided into two categories based on worker’s mobility patterns: (i) Some existing research works for MCS task assignment only consider opportunistic mode where workers will complete tasks assigned to them by MCS platform without changing their scheduled movement route in their daily life. In [5–8], the authors studied the assignment of single task in MCS. They proposed different task assignment schemes, which can maximize the sensing quality of tasks or assigned tasks to a minimum number of workers to ensure a certain level of tasks’ sensing quality. In contrast, some authors studied task assignment of multi-task coexistence in MCS, where tasks share limited resources such as budget and workers. For example, both [9] and [10] proposed multi-task assignment schemes to maximize overall system utility when the tasks share a limited incentive budget. (ii) The other category considers the task assignment under the participatory mode in MCS, where workers don’t have a fixed movement route and they will go to the designated location to perform the task. For instance, Kazemi et al. [11] aimed to maximize the number of completed tasks, while ensuring constraints on worker’s maximum number of accepted tasks. Similarly, Hu et al. [12] considered minimize the sensing cost for completing a given tasks set under the constraint on worker’s maximum movement distance. Most existing task assignment schemes adopt either the opportunistic mode or participatory mode while they did not consider task assignments in which two sensing modes coexisted.

As a matter of fact, both sensing modes have their own advantages and disadvantages, which we will explain as follow. The opportunistic mode does not require to change the worker’s scheduled movement route, so the interference to the worker is relatively small, and the cost of the task organizer is low. But the sensing quality of task largely depends on the worker’s movement route. For tasks occur in locations where are few workers pass through, the sensing quality may be very low. In participatory mode, workers can move according to sensing requirements of task, which can guarantee task completion. But there will be additional sensing cost to compensate the workers’ mobile overhead. Moreover, sensing tasks become complicated with the increase in demands of mobile crowdsensing, a task should be assigned to the workers who meet sensing requirements of tasks.

Motivated by the complementary nature of two sensing modes, we study the problem of the heterogeneous task assignment with the coexistence of two sensing modes. Compared to the task assignment problems in opportunistic or participatory mode alone, there are two research challenges in our problem. First, since the workers of two types (i.e., opportunistic workers and participatory workers) share a total budget, we need a more complex method for task assignment which considers the workers of two sensing modes simultaneously. Second, due to the heterogeneity, tasks should be assigned to workers who meet

their sensing requirements. In an effort to address the problem and challenges mentioned above, our work makes the following contributions:

- We formulate the problem of heterogeneous task assignment with the coexistence of two sensing modes, in which tasks have different requirements on sensing time, location and sensor.
- We propose a two-phase greedy algorithm MSHTA to address our problem. MSHTA aggregates the respective advantages of the two sensing modes and assign the task based on the respective requirements of task.
- We evaluate the performance of the MSHTA through extensive simulations. The experimental results demonstrate the effectiveness of our proposed algorithm in terms of tasks’ sensing quality and tasks’ sensing cost.

2 System Model and Problem Statement

In this section, we first introduce the system model of our MCS scenario, and then formulate the problem of the heterogeneous task assignment with the coexistence of two sensing modes.

2.1 System Model

We consider a heterogeneous task assignment problem in an MCS scenario where two sensing modes coexist. Specifically, the entire sensing region is divided into l sub-regions where the data collection mission in one sub-region is defined as a “sensing task”. Suppose that there are m sensing tasks during a certain time period where each task has its own sensing requirements (e.g., sensing time, sensing sensor), denoted by $\mathcal{T} = \{t_1, t_2, \dots, t_m\}$. For simplicity, we divide the certain time period into g equal slots. The task t_j ($j \in \mathcal{T}$) can be characterized by a parameter tuple (s_j, e_j, l_j, d_j) , where s_j is the start slot of the task t_j , e_j is the end slot of the task t_j , l_j represents the sensing region of the task t_j , d_j indicates the required sensor type of the task t_j . We assume that the union of sensors is denoted as $\mathcal{S} = \{d_1, d_2, \dots, d_q\}$. The energy consumed of task t_j is determined by the type of sensor and the sensing slots. Suppose that the energy consumed by all sensors each slot is known, represented by the set \mathcal{R} , where $\mathcal{R} = \{r_{d_1}, r_{d_2}, \dots, r_{d_q}\}$.

On the other hand, we assume there are n workers who have registered in the MCS platform and they are willing to collect the data which sensing tasks need. The workers can be divided into the opportunistic workers and participatory workers according to the willingness of workers. Denote $W_O = \{w_1, w_2, \dots, w_o\}$ the opportunistic worker set and $W_P = \{w_1, w_2, \dots, w_p\}$ the participatory worker set, respectively. We adopt a parameter tuple (l_i, s_i, k_i, b_i) to characterize the worker w_i ($w_i \in W_o \cup W_p$). Specifically, s_i is the set of sensors held by worker w_i , k_i is the operational proficiency set of the sensor that the worker w_i holds, it can be derived from the worker history execution task record, l_i is the current location of the worker w_i , and b_i is the current remaining battery of the mobile

device held by the worker w_i . Suppose that the location of the worker does not change during the slot but changes between slots.

We adopt a centralized task allocation model, in which MCS platform collects the worker's historical data and then selects the appropriate workers from the worker pool to perform the task [13,14]. Since the sensing quality of task depends on the opportunistic worker's movement route, it is necessary to accurately predict the location of opportunistic workers. A common method is that MCS platform uses the historical movement trajectories of the opportunistic worker to characterize the opportunistic worker's mobility. Considering the difference in tasks sensing time, we directly use a statistical-based model to derive the probability that an opportunistic worker will pass a particular region at a given period.

Denote \mathcal{D} the historical trajectory data set of the opportunistic worker, wherein each record consists of a worker ID , current location, and time period. Based on the learning of existing location records, the probability that the worker w_i will access the specific region r during time period t can be calculated as follow:

$$P(w_i, t, r) = \frac{|\mathcal{D}_{(w_i, t, r)}|}{|\mathcal{D}|}, \quad (1)$$

where $|\mathcal{D}_{(w_i, t, r)}|$ represents the number of days in which the worker w_i pass the region r during time period t throughout \mathcal{D} , and $|\mathcal{D}|$ is the number of days included in the data set \mathcal{D} .

According to the mobility of the participatory worker, the probability that a participatory worker appears in the task sensing region within the task sensing time is 1.

In addition, in order to ensure the sensing quality of task, each task should be executed by multiple workers and then MCS platform aggregates the returned data by workers to obtain a more realistic value. In this paper, γ is specified as the number of workers performing tasks. We use the binary variable a_{ij} to indicate whether or not the task t_j is assigned to the worker w_i . $a_{ij} = 1$ means that w_i is selected to perform task t_j , otherwise 0. Therefore, the task assignment solution can be formalized as follow:

$$a_{ij} = \begin{cases} 1 & \text{task } t_j \text{ is assigned to worker } w_i, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Furthermore, to ensure the sensing quality of the task, a worker can only perform one task at a slot.

When a worker goes to perform a task, it will lead to a cost, which is composed of three parts. The first part is basic reward paid to each worker recruited to the MCS platform. The second part is the sensing cost of workers using their mobile devices to perform sensing task. The third part is the cost of workers moving to the location of tasks. To compensate for the cost, the worker will receive a reward from the MCS platform. Specifically, if a worker w_i is assigned multiple tasks, w_i will receives more reward. Therefore, the reward for w_i can be expressed as:

$$R(w_i) = R_0 + R_1 * \sum_{j=1}^m a_{ij} * |e_j - s_j + 1| + R_2 * \sum_{j=1}^m a_{ij} * d(w_i, t_j), \quad (3)$$

where R_0 is basic reward paid to each worker who is recruited to the MCS platform; R_1 represents the reward that the worker performs the sensing task at each slot; R_2 represents the unit distance movement reward for the workers' movement. The distance between the worker and the task is defined as the number of sub-regions that the worker passes through when he arrives at the task sensing regions. Moreover, all tasks share a total budget B , that is, the reward paid to all workers cannot exceed B .

2.2 Problem Statement

We use the task's coverage ratio and workers' sensor operation proficiency to measure the task's sensing quality, where the coverage ratio of task is the probability that the worker passes the task's sensing location during the task's sensing time. Given an MCS task t_j , its sensing quality is calculated as follow:

$$Q_{t_j} = \frac{1}{\gamma} * \sum_{i=1}^n a_{ij} * P(w_i, l_j, t_{s_j \sim e_j}) * k_i, \quad (4)$$

where

$$P(w_i, l_j, t_{s_j \sim e_j}) = \frac{\sum_{t=s_j}^{e_j} P(w_i, l_j, t)}{|e_j - s_j + 1|}. \quad (5)$$

Then, we define the utility value formula to calculate the utility value that the worker w_i performs the task t_j as follow:

$$u_{w_i} = P(w_i, l_j, t_{s_j \sim e_j}) * k_i. \quad (6)$$

Based on the above description and assumption, the system objective is to maximize the overall sensing quality of tasks under the budget constraint:

$$\max_a \sum_{j=1}^m Q_{t_j} \quad (7)$$

s.t.

$$C_1 : \sum_{j=1}^m a_{ij} * |t_j| * b_{d_j} \leq b_i, \forall i \in W_o \cup W_p \quad (8)$$

$$C_2 : \sum_{i=1}^n a_{ij} = \gamma, \forall j \in T \quad (9)$$

$$C_3 : \sum_{i=1}^n R(w_i) \leq B \quad (10)$$

$$C_4 : s_j \subset s_i, \forall j \in T, \exists i \in W_o \cup W_p \quad (11)$$

$$C_5 : \sum_{j=1}^m a_{ij} \leq 1, \forall i \in W_o \cup W_p \quad (12)$$

Constraint C_1 means that each worker is assigned a task set whose consumed energy cannot exceed the worker's available battery. Constraint C_2 represents all tasks share the total budget B . Constraint C_3 represents each task needs to be performed by γ workers. Constraint C_4 represents if a worker is assigned a task, he must have the sensor required of the task. Constraint C_5 represents at most one task is assigned to a worker in a slot.

When we do not consider the heterogeneity of tasks, budget constraint and workers' battery constraint, this problem could be simplified into a Knapsack problem, which is NP-hard.

3 Problem-Solving Algorithms

Considering the time complexity and optimality of the algorithm, we design a two-phase greedy algorithm to solve our problem, namely MSHTA.

Note that the final achieved assignment may not be real optimal solution, but rather close to the real optimal solution. The underlying reason is that we do not optimize the search space globally but decompose the original search space into multiple subspaces and probe the local optimization solutions one by one. However, by this method, the computational overhead can be significantly reduced.

3.1 Task Optimization Allocation: Greedy-Based Search

In this section, we will explain the task assignment process. The task assignment process is mainly composed of two phases. In the first phase, the task is assigned to the opportunistic worker, and in the second phase, the tasks with lower sensing quality are reassigned to the participatory worker execution to improve the sensing quality of task. In the following, the workflow of the algorithm will be described.

In the first phase, for each task, we first remove the invalid opportunistic worker who does not pass the task sensing region (i.e., $P=0$) during the task execution time or does not have the sensor required for the task or with insufficient remaining battery to perform the task. Then, for each qualified opportunistic worker, the worker's utility value for the task is obtained based on the utility value formula, and the incentive increment generated by assigning the task to

the worker will be calculated by reward formula. Finally, a sorting operation is performed on all qualified workers based on the utility value of the worker. And select top- γ qualified workers as the performer of the task. The corresponding binary variable a_{ij} is set to 1. The selected worker list SeleList and the available budget will be update accordingly. The process continues until all tasks are associated with γ workers or the budget is exhausted or there are no available opportunistic workers. Related pseudocode is presented in Algorithm 1.

Algorithm 1. MSHTA:phase-1

```

1: Input : opportunistic worker set  $W_O$ , task set  $T$  and budget  $B$ 
2: Output : the binary variable  $a_{ij}$  and the selected worker list SeleList
3: for each task in  $T$  do
4:   while  $\sum_{i=1}^o R(w_i) < B$  do
5:     for each worker in  $W_O$  do
6:       if  $P(w_i, t_j, l_j)$  is zero or don't have corresponding sensor or  $\sum a_{ij} * |e_j - s_j + 1| * r_{d_j} \geq b_i$  then
7:         skip  $w_i$  and examine next worker in oppWorker
8:       end if
9:       calculate utility value and costs of  $w_i$ 
10:       $Rank_w = \text{Sorting}(w_i, 1 \leq i \leq o)$  based on utility value
11:      Choose top- $\gamma$  workers as the task performer
12:      Update seleList, budget and  $a_{ij}$ 
13:    end for
14:  end while
15: end for

```

In the second phase, all tasks are sorted based on the current sensing quality. In each iteration, we choose the lowest-sensing-quality task as the target. Participatory workers without the required sensor or with insufficient remaining battery are first removed, which can identify the qualified workers for the target task. Then for each qualified participatory worker, we obtain its utility based on the utility function, and the incentive increment generated by assigning the task to the worker will be calculated by reward function. Finally, a sorting process is conducted on all qualified participatory workers based on the utility value. The participatory worker with largest utility value is compared with the original γ executive workers in the utility value. If the utility value of this participatory worker is greater than the utility value of one worker in the original γ executive workers, select the participatory worker who with the largest utility value and lowest incentive increment as the performer of the task and then delete an original execution worker who have the smallest utility value of the task. And thus, sort all tasks based on current sensing quality, update the selected worker list SeleList and the available budget. The process continues until the budget is exhausted or there are no available participatory workers. The relevant pseudocode is presented in Algorithm 2.

After the above two-phase task assignment process, each task is assigned to the worker **who most responsive the task's sensing requirements**.

Algorithm 2. MSHTA:phase-2

```

1: Input : partocopatory worker set  $W_P$ , task set  $T$ , the binary variable  $a_{ij}$  and
   current available budget  $B$ 
2: Output : the binary variable  $a_{ij}$  and the selected worker list SeleList
3: sorting  $T$  based on the sensing quality of task
4: for each task in  $T$  do
5:   while  $\sum_{i=1}^p R(w_i) < B$  do
6:     for each worker in  $W_P$  do
7:       if don't have corresponding sensor or  $\sum a_{ij} * |e_j - s_j + 1| * r_{d_j} \geq b_i$  then
8:         skip  $w_i$  and examine next worker in opWorker
9:       end if
10:      calculate utility value and costs of  $w_i$ 
11:       $Rank_w = \text{Sorting}(w_i, 1 \leq i \leq p)$  based on utility value;
12:      if the maximum utility value more than original utility value then
13:        Choose workers with the maximum utility value and less costs than other
        workers
14:        Update seleList, budget and  $a_{ij}$ 
15:      end if
16:    end for
17:  end while
18: end for

```

4 Experimental Purposes and Baselines

The purpose of our simulation is to compare the performance of the MSHTA and other benchmark schemes in different situations (e.g., different task number, different budget). The performance indicators are the average task sensing quality, the number of tasks completed, total task sensing cost and total movement distance of the worker. Specifically, we have provide the following two benchmark schemes for comparative research.

OPP (Opportunity Mode Based Approach): This algorithm only uses opportunistic workers to maximize the sensing quality of all tasks while maintaining budget constraint. It spends all budget to choose opportunistic workers to perform task. Similar to [7], the OPP iteration selects the worker until the total budget is exhausted or the sensing quality of task no longer increases, and the selected workers will complete the task in their daily life. This benchmark scheme is used to test whether the MSHTA scheme is more efficient than the pure opportunity scheme in terms of tasks' sensing quality, tasks' sensing cost and task completed number.

PAR (Participation Mode Based Approach): This algorithm only uses participatory workers to maximize the sensing quality of all tasks while maintaining budget constraint. PAR spends all budget to choose the participatory workers and workers move to the specified location to perform tasks. Specifically, it uses a greedy-based algorithm to select workers to perform tasks. This benchmark approach is used to test whether the MSHTA scheme is more efficient than the

pure participation scheme in terms of tasks’ sensing quality, tasks’ sensing cost, movement distance of the worker and task completed number.

4.1 Experimental Setups

In our simulation, we assume that tasks and workers are uniform distributed in the 5KM*5KM region which is divided into 25 sub-regions, and all tasks will occur between 10 O’clock and 12 O’clock where this time period is divided into 8 slots. The available sensors of the worker are random and the worker’s sensor operation proficiency follows the normal distribution with a parameter of (0.8, 0.2). The remaining battery of the mobile device held by the worker follows a normal distribution with the parameter (70, 20). The cost of each worker recruited to the MCS platform is set to 2, the cost of performing the task is set to 0.5 each slot, and the cost per unit of movement distance of the participatory worker is set to 2.

4.2 Evaluation Results

In this section we will present the results of the simulation experiment. We can observe the impact of various parameters on the sensing quality of task, the number of tasks completed, the total sensing cost and the total movement distance of the worker. If the task’s sensing quality is not less than 0.6, we say that the task is completed.

Impact of Total Budget. We first investigate the impact of the budget on the performance. We set the number of tasks to 30, the number of workers to 80 and the γ value to 3. The budget is change from 300 to 900. The results are shown in Figs. 1 and 2.

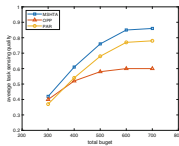


Fig. 1. Task sensing quality under different total budget.

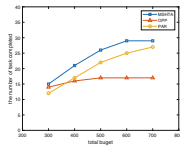


Fig. 2. Task completed number under different total budget.

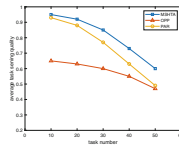


Fig. 3. Task sensing quality under different task number.

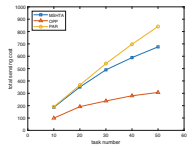


Fig. 4. Task sensing cost under different task number.

In Fig. 1, we can observe that, as the budget increases, the average task sensing quality also increases for three schemes. This is because a higher budget allows more workers recruited to perform tasks. Since our MSHTA scheme considers the advantages of the two sensing modes simultaneously, the average

sensing quality of our MSHTA scheme is more than other schemes. Moreover, our MSHTA scheme can complete more tasks than other two schemes, which is demonstrated by Fig. 2. Because the participatory workers in PAR need more incentives to perform tasks, it only can complete fewer tasks when the budget is 300, and the average task sensing quality in PAR is smallest.

Impact of Task Number. To study the impact of task number, we keep the number of workers, budget and γ value fixed, i.e., $n=80$, $B=600$ and $\gamma=3$, and task number in the range of 10, 20, 30, 40, 50. Then, we observe the changes in the task average sensing quality and sensing cost, as shown in Figs. 3 and 4.

It is obvious that in Fig. 3 the average task sensing quality decreases as the task number increasing, that is because with the number of task increases, more tasks require more workers and budget to perform tasks. Since MSHTA scheme aggregates the advantages of two sensing modes in terms of sensing cost and sensing quality, it has a higher average task sensing quality than other schemes.

Figure 4 shows the performance of three schemes in the sensing cost under different task number. Although the MSHTA scheme cannot achieve the lowest sensing cost due to the workers' mobile cost, but its task sensing quality is the highest among the three schemes.

Impact of γ Value. Then, we evaluate the impact of the γ value. The value of γ is varied from 1 to 5 with the increment of 1. We set the number of workers to 80, the budget to 600 and the number of tasks to 30. The results are shown in Figs. 5 and 6.

Figure 5 shows that as the value of γ increases, the average task sensing quality is decreasing, this is because as the γ increases, more worker and more budget are required to perform the task. Combined with Fig. 6, we can see that MSHTA scheme achieves the highest sensing quality of task when need a lower sensing cost. That is because MSHTA scheme assign tasks to workers with higher sensing quality and lower sensing cost.

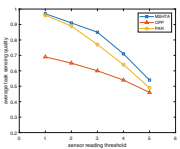


Fig. 5. Task sensing quality under different sensor reading threshold.

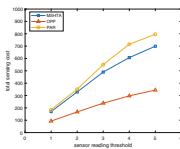


Fig. 6. Task sensing cost under different sensor reading threshold.

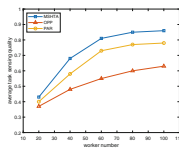


Fig. 7. Task sensing quality under different worker number.

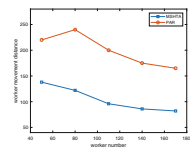


Fig. 8. Moving distance of worker under different worker number.

Impact of Worker Number. Finally, we study the impact of the worker's number on the performance. We set the number of tasks to 30, the γ value to 3 and the budget to 600. The results are shown in Figs. 7 and 8.

From Fig. 7, we can see that the average task sensing quality increase as the number of workers increases for three task allocation schemes. This is because as the number of workers increases, there will be workers who are more responsive for tasks' sensing requirements. And the performance of MSHTA scheme is better than other schemes under different number of workers.

As shown in Fig. 8, because the moving distance of the workers in our scheme is shorter than that in PAR scheme, our scheme can achieve a lower sensing cost. Compared with OPP scheme, although our scheme has a higher sensing cost, it is much higher than the OPP scheme in terms of tasks' sensing quality.

5 Conclusion

In this paper, we studied a novel heterogeneous task allocation problem in the MCS scenario where two sensing modes coexist. First, we introduced the system model of our MCS scenario and formalized the problem of the heterogeneous task assignment problem with the coexistence of two sensing modes. Then, we proposed a two-phase task assignment algorithm MSHTA to solve this problem, which aggregates the advantages of two sensing modes. In the first phase, MSHTA selects a set of opportunistic workers to perform tasks during their daily life; in the second phase, MSHTA reassigns the tasks with lower sensing quality to participatory workers and requires them move to the specified region to perform the task. Finally, simulation experiments show that the performance of our task assignment scheme is better than other two benchmark schemes.

Acknowledgment. This work is supported by the Natural Science Foundation of China (No. 61872104), the Natural Science Foundation of Heilongjiang Province in China (No. F2016009), the Fundamental Research Fund for the Central Universities in China (No. HEUCF180602) and the Tianjin Key Laboratory of Advanced Networking (TANK), College of Intelligence and Computing, Tianjin University, Tianjin China, 300350.

References

1. Ganti, R.K., Ye, F., Lei, H.: Mobile crowdsensing: current state and future challenges. *IEEE Commun. Mag.* **49**(11), 32–39 (2011)
2. Koukoumidis, E., Peh, L.-S., Martonosi, M.R.: 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 (2011)
3. Omokaro, O., Payton, J.: Flysensing: a case for crowdsensing in the air. In: 2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS), pp. 545–550. IEEE (2014)

4. Zappatore, M., Longo, A., Bochicchio, M.A., Zappatore, D., Morrone, A.A., De Mitri, G.: A crowdsensing approach for mobile learning in acoustics and noise monitoring. In: Proceedings of the 31st Annual ACM Symposium on Applied Computing, pp. 219–224. ACM (2016)
5. Reddy, S., Shilton, K., Burke, J., Estrin, D., Hansen, M., Srivastava, M.: Using context annotated mobility profiles to recruit data collectors in participatory sensing. In: Choudhury, T., Quigley, A., Strang, T., Sugiuma, K. (eds.) LoCA 2009. LNCS, vol. 5561, pp. 52–69. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-01721-6_4
6. Reddy, S., Estrin, D., Srivastava, M.: Recruitment framework for participatory sensing data collections. In: Floréen, P., Krüger, A., Spasojevic, M. (eds.) Pervasive 2010. LNCS, vol. 6030, pp. 138–155. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12654-3_9
7. Zhang, D., Xiong, H., Wang, L., Chen, G.: Crowdrecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 703–714. ACM (2014)
8. Wang, L., Zhang, D., Wang, Y., Chen, C., Han, X., M’hamed, A.: Sparse mobile crowdsensing: challenges and opportunities. *IEEE Commun. Mag.* **54**(7), 161–167 (2016)
9. Song, Z., Liu, C.H., Wu, J., Ma, J., Wang, W.: QoI-aware multitask-oriented dynamic participant selection with budget constraints. *IEEE Trans. Veh. Technol.* **63**(9), 4618–4632 (2014)
10. Wang, J., et al.: Fine-grained multitask allocation for participatory sensing with a shared budget. *IEEE Internet Things J.* **3**(6), 1395–1405 (2016)
11. Kazemi, L., Shahabi, C.: Geocrowd: enabling query answering with spatial crowdsourcing. In: Proceedings of the 20th International Conference on Advances in Geographic Information Systems, pp. 189–198. ACM (2012)
12. Hu, T., Xiao, M., Hu, C., Gao, G., Wang, B.: A QoS-sensitive task assignment algorithm for mobile crowdsensing. *Pervasive Mob. Comput.* **41**, 333–342 (2017)
13. Wang, L., Zhiwen, Y., Guo, B., Yi, F., Xiong, F.: Mobile crowd sensing task optimal allocation: a mobility pattern matching perspective. *Front. Comput. Sci.* **12**(2), 231–244 (2018)
14. Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., Han, J.: Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In: Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, pp. 1187–1198. ACM (2014)



A Lightweight Neural Network Localization Algorithm for Structureless Wireless Sensor Networks

Rong Gao, Zhongheng Yang, and Hejun Wu^(✉) 

Guangdong Key Laboratory of Big Data Analysis and Processing,
Department of Computer Science, Sun Yat-sen University, Guangzhou, China
wuhejun@mail.sysu.edu.cn

Abstract. This paper studies distributed range-based localization in arbitrarily deployed wireless ad hoc networks. Existing range-based localization approaches depend on specially deployed anchors or require dense network deployment. Our algorithm is a distributed paradigm that only requires local information of each node. Therefore, it is applicable to the resource-limited embedded sensors. Specifically, our algorithm performs a three-stage optimization through coarse-grained, middle-grained, and fine-grained levels. We designed an efficient but accurate neural network to learn the hidden relations between the distances of nodes and their positions. Simulations show that our proposed algorithm works in many more types of network deployments than the existing approaches. Furthermore, our algorithm achieves the highest localization accuracy on average.

Keywords: WANET · Range-based localization · Neural networks

1 Introduction

Multi-hop wireless ad hoc networks (WANETs) such as wireless sensor networks (WSNs) or wireless mesh networks (WMNs) are promising in various spot monitoring applications, especially in emergency scenarios. These ad hoc networks can be applied to predict or detect disasters and faults [1], such as underground mine monitoring [2] and underwater searching [3,4]. In such applications, the node location information is a prerequisite for locating where the interested event occurs. However, the nodes are usually randomly deployed in the monitored area due to limitations of space, strict application requirements, or large scales. As a result, localization is indispensable for randomly deployed multi-hop wireless networks. WANETs also apply to body monitoring in healthcare systems [5–7].

Two kinds of localization methods have been proposed: range-free and range-based schemes. The performance of a range-based method is usually better than that of a range-free method. Range-based localization methods include multidimensional scaling (MDS) [9–11], gradient descent based methods [12],

DV-distance [13] and rigidity-based methods [14–16]. Nonetheless, the existing range-based methods are limited in that they require special network topology structures or need the network to be dense.

To address this problem, we propose a localization algorithm that does not require carefully selected anchors or dense node deployments, named LNN. It is a distributed lightweight localization method with a neural network utilizing both distance information and hop information. In each iteration, each node uses a neural network to estimate its coordinate using coordinates of other nodes and its shortest paths to them (Some nodes cannot communicate with each other directly with one hop). Estimated distances to other nodes can be calculated using the estimated coordinate. A loss function is defined by these estimated distances and shortest paths and is used to train the LNN. Therefore, LNN learns the hidden relations between coordinates and distances of other nodes and the coordinate of the node itself. Finally every node broadcast its estimated coordinate to its neighbors to start the next training iteration. After several iterations, the estimated coordinates of all the nodes converge to their real coordinates.

The main contribution of this paper is the three-stage distributed algorithm to determine both coarse-grained positions and fine-grained positions of nodes in the network. For each node, the three stages are: (1) using positions of three anchors and distance to them set its initial estimated coordinate. (2) using neighbors within three hops to update the estimated coordinate of itself (3) using only direct neighbors to update the estimated coordinate of itself. With these stages to determine coarse-grained, mid-grained and fine-grained coordinates gradually, much fewer anchors in localization (only three in two-dimensional environments) are required. Due to the three-stage design of LNN, the convergence is quite fast. Since LNN is a distributed algorithm, each node only uses limited nodes to optimize its coordinate. A center node with high computing power is not required. Therefore, our algorithm applies to resource-limited embedded sensors and scalable networks. In addition, LNN's advantage on sparse deployments reduces the needed number of sensors to cover the same area than existing range-based algorithms. These advantages lead to both lower economic and lower energy costs. Simulations show that LNN significantly improves localization performance in terms of accuracy and efficiency.

2 Related Work

Localization methods can be classified as range-based methods and range-free methods [17]. The state-of-the-art performance of the range-free methods is achieved by a hybrid method integrated with the approximate point in triangulation (APIT) and distance vector-hop (DV-HOP) [18]. APIT is a method based on the PIT test. Given an unknown node, PIT can identify all the triangulated areas containing this node. The main idea of APIT is to compute the overlapping area and consider the centroid of this area as an estimation of the node's coordinate. DV-HOP is a method based on trilateration. In DV-HOP, the distances between each unknown node and each anchor are estimated using hop

information. The main idea of the DV-HOP algorithm is to compute the coordinate of each unknown node using the estimated distances from the anchors. These range-free approaches do not consider distance measurements and cannot achieve such a good performance as range-based approaches. Thus, we do not discuss these in detail.

Localization techniques with distance measurements can perform better than range-free techniques. There are measurement techniques in general, known as AOA measurements, RSS profiling measurements, and distance-related measurements [19]. AOA is short for angle-of-arrival, which makes use of either the amplitude or the phase response of the receiver antenna. Distance-related measurements can be further classified as one-way propagation, roundtrip propagation, lighthouse approach [20], RSS (received signal strength)-based measurements and TDOA (time-difference-of-arrival) measurements. RSS profiling measurements have been widely applied in WLAN localization and have begun to show their effect in other localization scenarios. [8, 21, 22] proposed methods to determine distances using signal energy. In [8], RSSI with 20-m range precision can reach 1.5 m. Different localization methods based on different measurements have been developed, with different accuracies and costs.

The DV-distance algorithm [13] is an advanced version of DV-HOP, which is used in this paper as an initial procedure. The difference between DV-distance and DV-HOP is the usage of distance information. Therefore, DV-distance can be regarded as a distance-included version of DV-HOP.

Multidimensional scaling (MDS) also has two versions, a connectivity-based version [11] and a distance-based version [10]. In terms of distance-based MDS, a centralized version is presented in [10], but a distributed implementation is also possible. The goal of MDS is to divide the network into groups. The relative positions of nodes in one group can be calculated with multidimensional scaling [23]. Then, the relative positions are aligned with the aid of anchors. Finally, the groups are combined, and the absolute positions are produced.

Shan et al. [24] proposed an approach for indoor localization using RSS. Measurements of distances were taken several times and calculated by the weighted average of these times. Refined distances were used for localization. The anchors were divided into many combinations of three. User localization was the mean of those estimated by all three anchor groups. In LNN, the first phase used mean positions estimated by different usages of anchors. Terán [25] proposed an indoor localization approach using a machine learning perspective. The K-NN search algorithm and K-means clustering algorithm were applied to determine the position of the node. Likewise, Margolies [26] proposed an offline-training and online-predicting scheme to determine user localization in a 4G LET network using different signal measurements.

Studies on rigidity show that determining relative localization can be performed only using relative distance measurements. Yang and Liu proposed a theoretical approach to detect localizable nodes using rigidity [14]. For localizable nodes, rigidity-based algorithms can determine their accurate positions. However, the performance of these algorithms depends highly on the topologies

of the networks and the selection of anchors. The reason is that the selection of anchors highly affects the ratio of localizable nodes in a network.

Gradient-based algorithms have been proposed to gain fine-grained localization accuracy. Savvides et al. [27] proposed a collaborative and distributed model to estimate nodes' locations using known anchor locations that are several hops away and distance measurements to neighboring nodes. The loss functions of this problem may be varied, but they have the same goal of reducing the distance error. Grag et al. [28] proposed a function of the difference between the estimated distances and the ground truth distances and calculated the derivative of this function to apply the gradient descent method.

Nguyen et al. [29] proposed a centralized gradient-based approach to determine the positions of nodes by reconstructing the distance matrix. Jie Cheng [30] proposed a distributed localization scheme based on distance matrix reconstruction and convex optimization named DISCO. DISCO uses several minimization problems that only involve convex optimization. Thus, DISCO can achieve high localization precision with low computation complexity. D. Qiao et al. [12] proposed two gradient descent schemes to achieve fine-grained performance. Gradient descent method A (GDA) expresses the coordinate of one node as the linear combination of distances related to it. For example, given a node with coordinate (x, y) and distances d_1, d_2, \dots, d_n , (x, y) is expressed as $x = w_1d_1 + w_2d_2 + \dots + w_nd_n$, and $y = v_1d_1 + v_2d_2 + \dots + v_nd_n$, where w_i and v_i are weights. Since distances are given, the distance error can also be represented by the weights.

The gradient descent operation is applied to the weights. Gradient descent method B (GDB) is more straightforward, as the gradient descent operation is applied to the coordinates instead of the weights. Qiao et al. [12] showed that GDA and GDB have the same performance. The LNN in this paper is also gradient-based but adapts neural networks to learn more hidden information, which is one of the main contributions of our work.

3 Algorithm

The main idea of the algorithm is to optimize the coordinates of the nodes iteration by iteration. In each iteration, each node utilizes the information about others, including their coordinates, distances, and hops to them. All the information is used as input for a neural network to estimate the coordinate of the node itself. Then this node calculates its distances to other node using its estimated coordinate and calculates the mean-squared error. This mean-squared error measures the differences between the estimated distances and real distances represented by the shortest paths. Via minimizing this mean-squared error, the neural network learns the hidden relations between information about other nodes and the coordinate of the node itself. After training the neural network, each node broadcasts its coordinate to its neighbors and starts the next iteration. Finally, the coordinates of all the nodes will converge to their real positions. Localization may be trapped into local optimums in which nodes have right position relations

to their nodes but have a wrong global positions. We provide the information of nodes within three hops to each node to determine its global position.

Table 1. Notations

Notations	Description
$d_{i,j}$	Real distance from node i to node j
$d_{i,j}^*$	Estimated distance from node i to node j
(x_i, y_i)	Real coordinate of node i
(x_i^*, y_i^*)	Estimated coordinate of node i
n	Number of pairs of nodes
m	Number of nodes
θ	Parameters in the neural network
$h_{i,j}$	Hops from node i to node j
N_i	Node set estimating node i 's coordinate
D_i	Real distances between node i and N_i
C_i^*	Estimated coordinates of nodes in N_i
X_i	X-coordinates of nodes in N_i
Y_i	Y-coordinates of nodes in N_i
H_i	Hops from node i to nodes in N_i

The notations are listed in Table 1. The real coordinate of node i is denoted (x_i, y_i) . The estimated coordinate of node i is denoted (x_i^*, y_i^*) . For each pair of interconnected nodes (e.g., node i and node j), the real distance, denoted $d_{i,j}$, is measured. For each pair of non-interconnected nodes, the distance is the shortest path in the graph between them. The estimated distance can be calculated using the estimated coordinate, as

$$d_{i,j}^* = \sqrt{(x_i^* - x_j^*)^2 + (y_i^* - y_j^*)^2} \quad (1)$$

In our algorithm, (x_i^*, y_i^*) is estimated via a neural network. For utilizing as much information as possible, the input of the neural network of one node includes the estimated coordinates of nodes nearby node i (those nodes that can communicate with node i within a certain hop) and distances to them. The neural network is constructed with two hidden layers using sigmoid as the activation function. In other words, the neural network can be regarded as a nonlinear function:

$$(x_i^*, y_i^*) = f(\theta; D_i, X_i, Y_i) \quad (2)$$

where θ denotes all the trainable parameters of the neural network.

There are two types of loss functions, as illustrated in the following two formulas (L_1 and L_2). The first one is the real loss of coordinates

$$L_1 = \sum_i \sqrt{(x_i^* - x_i)^2 + (y_i^* - y_i)^2} / m \quad (3)$$

where m is the number of nodes.

Since LNN is a distributed algorithm, nodes do not know their real coordinates. They cannot use L_1 to train their neural networks. We need another applicable loss function. The second one is the distance error:

$$L_2 = \sum_{i,j} (d_{i,j}^* - d_{i,j})^2 / n \quad (4)$$

where n is the number of pairs of nodes.

Generally, L_2 is used for training, and L_1 is used for testing. We need to address the situation in which L_2 is low but L_1 is high, which may be caused by rotations and symmetries of some nodes to other nodes. We consider the nodes within more than one hop together with those with one hop and give them lower weights. We can use a *discount* factor to perform these weights. Then, we obtain L_3 :

$$L_3 = \sum_{i,j} \text{discount}^{h_{i,j}-1} (d_{i,j}^* - d_{i,j})^2 / n \quad (5)$$

Since L_3 is still a global loss function, in our distributed algorithm, each node needs to minimize its own part of L_3 . Then, we obtain the loss function to minimize for each node. For node i , we have the following loss function:

$$L = \sum_j \text{discount}^{h_{i,j}-1} (d_{i,j}^* - d_{i,j})^2 / |N_i| \quad (6)$$

where $|N_i|$ denotes the number of nodes used to estimate node i 's coordinate and j indexes these nodes.

Since we have a loss function, we can use the gradient descent method to minimize it and optimize the neural network as follows:

$$\theta_{new} = \theta - \frac{\partial L}{\partial \theta} \quad (7)$$

Then, we use θ_{new} to calculate the newly estimated $(x_i^*, y_i^*)_{new}$

$$(x_i^*, y_i^*)_{new} = f(\theta_{new}; D_i, X_i, Y_i) \quad (8)$$

After estimating the coordinate, the node will broadcast the new coordinate. Thus, the coordinates of all nodes can be optimized iteratively.

Multi-hop distances are more useful for determining a node's global coarse-grained position, while one-hop distances are more useful for finding its precise position. Thus, we divide our algorithm into three stages. In each stage, different nodes are used to optimize the coordinate of the node.

3.1 Pre-training

Obviously, if one node can directly communicate with all the anchors, we can use a geometric method to specify its coordinate. Using two anchors, we can find

two possible positions for one node. Then, using the last anchor, we can select the correct one from the two candidates.

If this node cannot communicate directly with all of the anchors, we can also use the inaccurate distances to find an approximate coordinate for the node. This operation gives us a better start for our algorithms than setting all the nodes at random. This preprocessing procedure is proposed as the DV-distance.

3.2 Training

After initializing the coordinate of nodes, we can use the information of nodes within a given hop to estimate a node's coordinate. In each iteration, for each node, we train the network to minimize its loss function. After training the network, we broadcast the new coordinate of this node to other nodes and go to the next iteration. When a node's coordinate does not change substantially in an iteration, the algorithm on this node will go on to the next stage, fine-tuning.

3.3 Fine-Tuning

After several iterations (about 5–10), each node will have an acceptable coordinate. Information about remote nodes can become noise that prevents the algorithm from reaching a lower global loss. Therefore, each node should modify its loss function and only consider one-hop neighbors and use native gradient descent to perform fine-tuning. After tuning, the node will also broadcast its new coordinate to the other nodes.

We will show the necessities of three stages in the section Simulation. Algorithm 1 shows the distributed procedure of LNN on each node.

4 Simulation

We simulated and evaluated our algorithm on randomly generated data. We generated one thousand sensor networks, each of which had twenty nodes, including three anchors with known coordinates.

In each simulation, we chose an image and three nodes as anchors at random. For different position relations of anchors, the algorithm showed different performances. Therefore, we evaluated our algorithm within different limitations on anchors.

To compare different algorithms, we set 3 criteria. An intuitive score was the average of the nodes' Euclidean distances between their real and estimated positions. Furthermore, we wanted to see if the algorithm could determine the global information of the sensor networks. We grouped the experiment results by their mean errors and compared the distributions of them of different algorithms. Error reduction curves were also provided to show the speed of convergence.

Algorithm 1. Algorithm on node i

Require:Coordinates of anchors, N_i , D_i , H_i **Ensure:** (x_i, y_i)

- 1: Pre-train: use coordinates of anchors and distances to them to initialize node i 's coordinates, as (x_i^*, y_i^*)
 - 2: Broadcast the initial coordinate to N_i , use received coordinates as C_i^*
 - 3: **while** (x_i^*, y_i^*) is changing **do**
 - 4: Update C_i^* as received coordinates
 - 5: Update (x_i^*, y_i^*) as $f(\theta; D_i, X_i, Y_i)$
 - 6: Update θ as $\theta - \frac{\partial L}{\partial \theta}$
 - 7: Broadcast (x_i^*, y_i^*) to N_i
 - 8: Use one-hop nodes as N_i , update D_i , H_i and C_i^*
 - 9: **for** $t = 1$ to T **do**
 - 10: // T is the iteration times set for fine-tuning
 - 11: Update θ as $\theta - \frac{\partial L}{\partial \theta}$
 - 12: Update (x_i^*, y_i^*) as $f(\theta; D_i, X_i, Y_i)$
 - 13: Broadcast (x_i^*, y_i^*) to N_i
 - 14: Update C_i^* as received coordinates
 - 15: **return** (x_i^*, y_i^*)
-

4.1 Simulation Setup

In simulations, we will not focus on how nodes communicate with each other and how information is transmitted in the network. At the very beginning, each node knows lengths of its shortest paths to other nodes and the abstract coordinates of the three anchors.

In each simulation, 20 nodes are randomly scattered on a range of $10 * 10$. Each node has a communication range of 3 or 4 (both settings will be tested). We use r to represent the communication range. For the choice of anchors, we only have two limitations.

A. Anchors should not be too close to each other.

B. Three anchors should not be on the same line.

Suppose we have three anchors. For limitation A, we ensure that the three anchors cannot directly communicate with each other. We use g_1 , g_2 and g_3 to represent the gradients of the three lines generated by the three nodes. For limitation B, we add a limit to g_1 , g_2 and g_3 to ensure that the three anchors are scattered as a triangle instead of a line. We ensure that

$$\left| 1 - \left| \frac{g_i}{g_j} \right| \right| > l (i \neq j) \quad (9)$$

where l is the limit we set for the gradients. A higher l means a stricter limitation on the position relationships of the anchors.

To simulate that the algorithm is running in a distributed environment, in each iteration, we optimize all nodes' coordinates one time in a random sequence.

For the noise of distance measurements, each distance between two nodes is multiplied by a random number ranging from 0.95 to 1.05.

4.2 Localization Samples

Figures 1 and 2 show four methods on two sensor networks with and without noises, respectively (Noises may occur for distance measurement). Points represent the real positions of the nodes, and crosses represent the estimated positions. Three inverted triangles represent three anchors. If two nodes can directly communicate with each other, a line is drawn between them. We show the differences between the estimated results using LNN, GDA, DV-distance, and MDS-MAP in noisy and non-noisy environments.

As shown in Figs. 1 and 2, LNN can handle scenarios that others cannot. In these situations, GDA shows the most similar results to LNN. However, a few nodes estimate a coordinate that is far from its real position. For DV-distance and MDS-MAP, nodes near the anchors have acceptable estimated coordinates, while others do not. In settings with noise, for the nodes, GDA gives an incorrect position, and LNN shows acceptable performance.

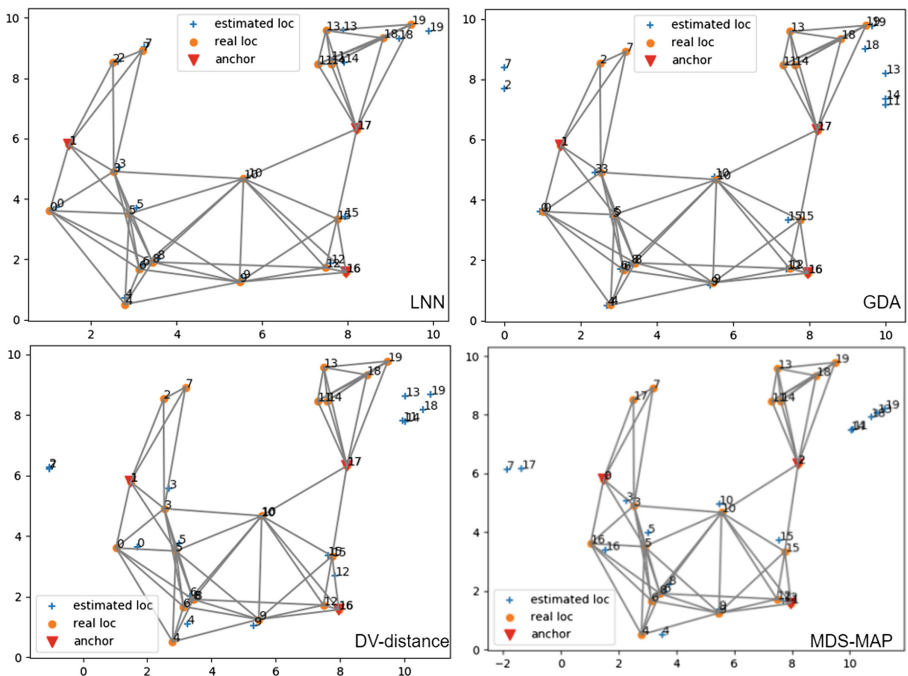


Fig. 1. Estimation results of four algorithms. Points are the real positions of nodes, crosses are estimated positions, and inverted triangles are anchors. Simulations are performed using 3 m as the communication range with 5% noise.

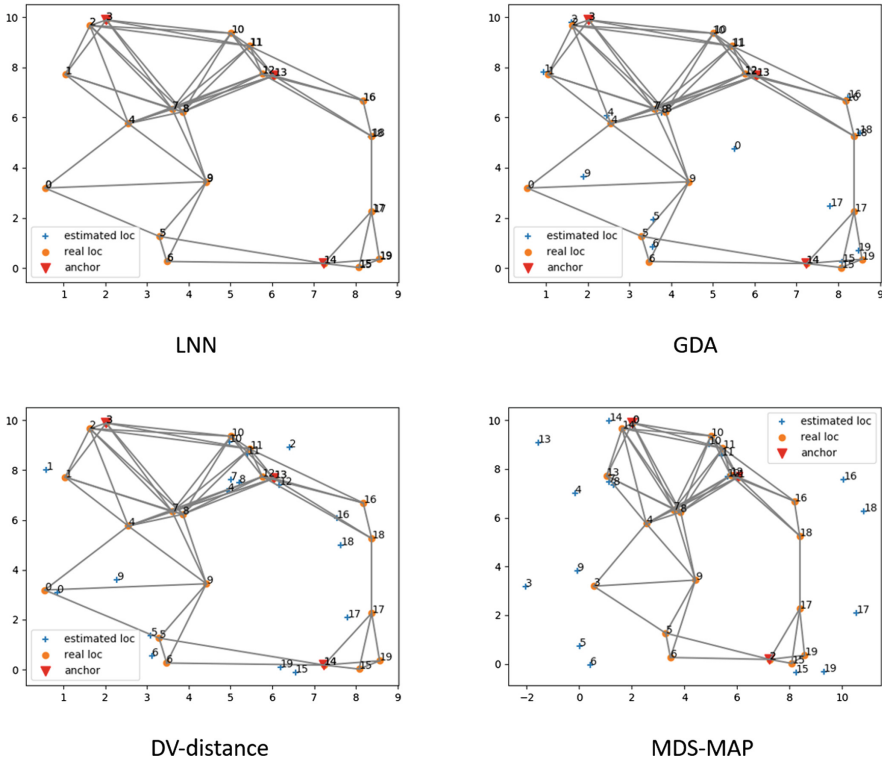


Fig. 2. Estimation results of four algorithms. Points are the real positions of the nodes, crosses are estimated positions, and inverted triangles are anchors. In the simulation, the wireless communication range is four meters with no noise.

4.3 Error Reduction Curve

Figures 3 and 4 show the performance of LNN in comparison with GDA. Figures 5 and 6 show the necessity of all three phases in our algorithm. Each curve shows the average performance of a simulation setting with one thousand simulations.

Figures 3 and 4 show the error reduction curves of LNN and GDA. The Y-axis is the average of the mean squared error (MSE) of the nodes' estimated coordinates against their real positions. The X-axis is the number of iterations. As shown in the two figures, the two algorithms show different performances under different simulation settings. LNN has a more obvious advantage in the short communication range with noise. Two algorithms using other settings show similar performances as the two simulation results shown in Figs. 3 and 4. In general, LNN needs fewer iterations to converge and has better accuracy than GDA to different degrees.

We also perform some simulations to show the necessity of the three phases in our algorithm.

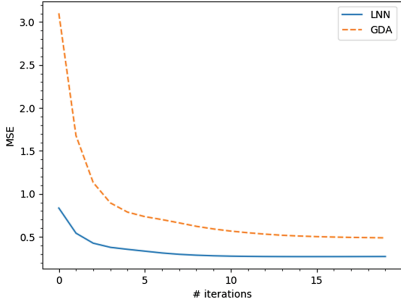


Fig. 3. Error reduction curve, $r = 3$, $l = 0.5$, with noise.

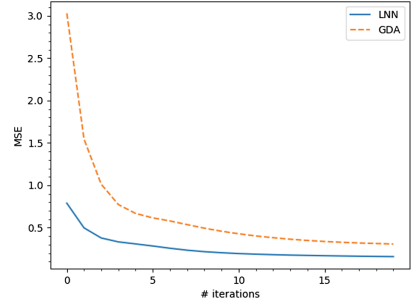


Fig. 4. Error reduction curve, $r = 4$, $l = 0.5$, with no noise.

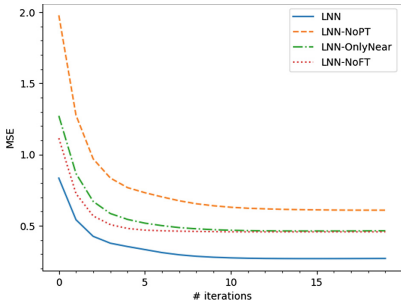


Fig. 5. Error reduction curve, $r = 3$, $l = 0.5$, with noise.

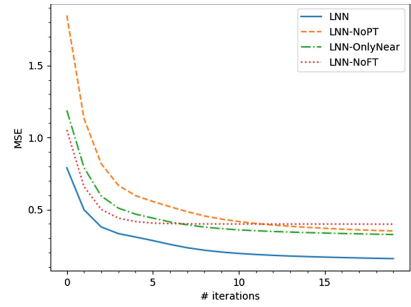


Fig. 6. Error reduction curve, $r = 4$, $l = 0.5$, with no noise.

As shown in Figs. 5 and 6, the three phases are important for localizing a node. LNN-NoPT means that all coordinates of the nodes are initialized at random in the space instead of using DV-distance initially. LNN-OnlyNear means that in the training phase, we only use the coordinates of neighbors and distances to neighbors to train the neural network. LNN-NoFT means we do not apply a fine-tuning phase after training and continue training for additional iterations. If fine-tuning is not applied, the inaccuracy of multi-hop distances prevents the MSE from decreasing to a lower level. In the other two situations, the algorithm can be trapped in local optimums. Under other settings, algorithms show similar performances as the two simulation settings shown in Figs. 5 and 6.

4.4 Error Distributions

We analyze the distributions of MSE of different algorithms, including LNN, GDA, DV-distance, and MDS-MAP, under different simulation settings.

As shown in Figs. 7 and 8, LNN has the most samples with an MSE under 0.3. It shows that LNN has the best ability to avoid local optimums. In addition, the communication range does not considerably affect the performance of our algorithm, which is different from GDA. DV-distance and MDS-MAP have

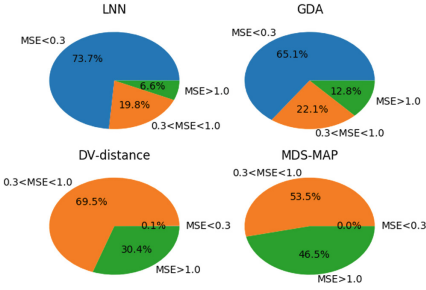


Fig. 7. MSE distributions, $r = 3$, $l = 0.5$, with noise.

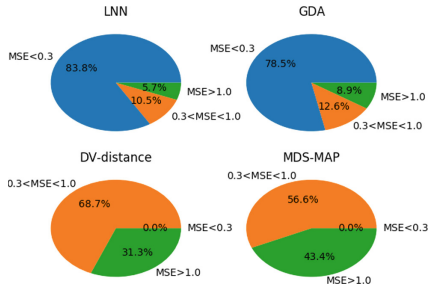


Fig. 8. MSE distributions, $r = 4$, $l = 0.5$, with no noise.

acceptable coarse-grained MSEs but cannot determine an accurate location for every node.

4.5 Comparison of Algorithms

As shown in Table 2, LNN has better accuracy than other algorithms to different degrees. Simulation results also show that LNN is not as sensitive to the communication range as GDA. For example, when the communication range decreases from 4 to 3, the MSE of LNN increases from 0.3 to 0.33, while the MSE of GDA increases from 0.36 to 0.48. It means that LNN can do better in sparse networks. For noise, LNN also shows better robustness than GDA. DV-distance and MDS-MAP are not sensitive to changes in settings. However, the nodes' locations cannot be accurately estimated.

Table 2. MSEs of different algorithms

Setting	LNN	GDA	DV-distance	MDS-MAP
With noise				
$r = 3$ $l = 0.5$	0.33	0.48	1.01	1.08
$r = 3$ $l = 0.2$	0.36	0.51	1.07	1.11
$r = 4$ $l = 0.5$	0.30	0.36	1.04	1.05
$r = 4$ $l = 0.2$	0.28	0.38	1.03	1.06
With no noise				
$r = 4$ $l = 0.5$	0.19	0.27	1.05	1.04
$r = 4$ $l = 0.2$	0.20	0.28	1.06	1.05

5 Conclusion

In this paper, a new localization method is proposed. Compared with pure gradient descent, the appended neural network architecture can learn much more hidden information about the location relations of nodes. In general, the proposed

method achieves a higher localization accuracy and gains better robustness, as shown in the simulation. Strict limitations on the topology of the network and the careful selection of anchors are not required. Future work is focused on the weights used in the loss function to further reduce the probability of local optima. An adaptive weighted approach may be a direction for exploration.

References

1. Othman, M.F., Shazali, K.: Wireless sensor network applications: a study in environment monitoring system. *Procedia Eng.* **41**, 1204–1210 (2012)
2. Minhas, U.I., Naqvi, I.H., Qaisar, S., Ali, K., Shahid, S., Aslam, M.A.: A WSN for monitoring and event reporting in underground mine environments. *IEEE Syst. J.* **12**(1), 485–496 (2018)
3. Sandeep, D., Kumar, V.: Review on clustering, coverage and connectivity in underwater wireless sensor networks: a communication techniques perspective. *IEEE Access* **5**, 11176–11199 (2017)
4. Zhou, F., Li, Y., Wu, H., Ding, Z., Li, X.: ProLo: localization via projection for three-dimensional mobile underwater sensor networks. *Sensors* **19**(6), 1414 (2019). <https://doi.org/10.3390/s19061414>
5. Pirbhulal, S., Zhang, H., Wu, W., Mukhopadhyay, S.C., Zhang, Y.: Heart-beats based biometric random binary sequences generation to secure wireless body sensor networks. *IEEE Trans. Biomed. Eng.*, 1 (2018). <https://doi.org/10.1109/TBME.2018.2815155>
6. Wu, W., Zhang, H., Pirbhulal, S., Mukhopadhyay, S.C., Zhang, Y.: Assessment of biofeedback training for emotion management through wearable textile physiological monitoring system. *IEEE Sensors J.* **15**(12), 7087–7095 (2015). <https://doi.org/10.1109/JSEN.2015.2470638>
7. Wu, W., Pirbhulal, S., Zhang, H., Mukhopadhyay, S.C.: Quantitative assessment for self-tracking of acute stress based on triangulation principle in a wearable sensor system. *IEEE J. Biomed. Health Inform.*, 1 (2018). <https://doi.org/10.1109/JBHI.2018.2832069>
8. Jie, Z., HongLi, L., et al.: Research on ranging accuracy based on RSSI of wireless sensor network. In: 2010 2nd International Conference on Information Science and Engineering (ICISE), pp. 2338–2341. IEEE (2010)
9. Shang, Y., Ruml, W., Zhang, Y., Fromherz, M.P.J.: Localization from mere connectivity. In: Proceedings of the 4th ACM International Symposium on Mobile Ad Hoc Networking and Computing, MobiHoc 2003, pp. 201–212. ACM, New York (2003). <https://doi.org/10.1145/778415.778439>
10. Ji, X., Zha, H.: Sensor positioning in wireless ad-hoc sensor networks using multidimensional scaling. In: IEEE INFOCOM 2004, vol. 4, pp. 2652–2661 (2004). <https://doi.org/10.1109/INFCOM.2004.1354684>
11. Shang, Y., Rumi, W., Zhang, Y., Fromherz, M.: Localization from connectivity in sensor networks. *IEEE Trans. Parallel Distrib. Syst.* **15**(11), 961–974 (2004)
12. Qiao, D., Pang, G.K.: Localization in wireless sensor networks with gradient descent. In: IEEE Pacific Rim Conference on Communications, Computers and Signal Processing Conference Proceedings. IEEE (2011). The Journal's web site is located at <http://www.ieeexplore.ieee.org/xpl/conhome.jsp?punumber=1000106>
13. Niculescu, D., Nath, B.: Ad hoc positioning system (APS). In: 2001 IEEE Global Telecommunications Conference, GLOBECOM 2001, vol. 5, pp. 2926–2931. IEEE (2001)

14. Yang, Z., Liu, Y.: Understanding node localizability of wireless ad hoc and sensor networks. *IEEE Trans. Mob. Comput.* **11**(8), 1249–1260 (2012)
15. Wu, H., Ding, Z., Cao, J.: GROLO: realistic range-based localization for mobile IoTs through global rigidity. *IEEE Internet Things J.*, 1 (2019). <https://doi.org/10.1109/JIOT.2019.2895127>
16. Wu, H., Ding, A., Liu, W., Li, L., Yang, Z.: Triangle extension: efficient localizability detection in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **16**(11), 7419–7431 (2017). <https://doi.org/10.1109/TWC.2017.2748563>
17. Dil, B., Dulman, S., Havinga, P.: Range-based localization in mobile sensor networks. In: Römer, K., Karl, H., Mattern, F. (eds.) *EWSN 2006*. LNCS, vol. 3868, pp. 164–179. Springer, Heidelberg (2006). https://doi.org/10.1007/11669463_14
18. Liu, C., Liu, S., Zhang, W., Zhao, D.: The performance evaluation of hybrid localization algorithm in wireless sensor networks. *Mob. Netw. Appl.* **21**(6), 994–1001 (2016)
19. Mao, G., Fidan, B., Anderson, B.D.: Wireless sensor network localization techniques. *Comput. Netw.* **51**(10), 2529–2553 (2007)
20. Römer, K.: The lighthouse location system for smart dust. In: *Proceedings of the 1st International Conference on Mobile Systems, Applications and Services*, pp. 15–30. ACM (2003)
21. Li, Z., Xiao, F., Wang, S., Pei, T., Li, J.: Achievable rate maximization for cognitive hybrid satellite-terrestrial networks with AF-relays. *IEEE J. Sel. Areas Commun.* **36**(2), 304–313 (2018)
22. Li, Z., Chang, B., Wang, S., Liu, A., Zeng, F., Luo, G.: Dynamic compressive wide-band spectrum sensing based on channel energy reconstruction in cognitive internet of things. *IEEE Trans. Ind. Inform.* **PP**(99), 1 (2018)
23. Borg, I., Groenen, P.: *Modern multidimensional scaling: theory and applications*. *J. Educ. Meas.* **40**(3), 277–280 (2003)
24. Shan, G., Park, B.-H., Nam, S.-H., Kim, B., Roh, B.-H., Ko, Y.-B.: A 3-dimensional triangulation scheme to improve the accuracy of indoor localization for IoT services. In: *2015 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pp. 359–363. IEEE (2015)
25. Terán, M., Aranda, J., Carrillo, H., Mendez, D., Parra, C.: IoT-based system for indoor location using Bluetooth low energy. In: *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*, pp. 1–6. IEEE (2017)
26. Margolies, R., et al.: Can you find me now? Evaluation of network-based localization in a 4G LTE network. In: *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9. IEEE (2017)
27. Savvides, A., Park, H., Srivastava, M.B.: The bits and flops of the N-hop multilateration primitive for node localization problems. In: *Proceedings of the 1st ACM International Workshop on Wireless Sensor Networks and Applications*, pp. 112–121. ACM (2002)
28. Garg, R., Varna, A.L., Wu, M.: Gradient descent approach for secure localization in resource constrained wireless sensor networks. In: *2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pp. 1854–1857. IEEE (2010)
29. Nguyen, L., Kim, S., Shim, B.: Localization in internet of things network: matrix completion approach. In: *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–4. IEEE (2016)
30. Cheng, J., Ye, Q., Du, H., Liu, C.: DISCO: a distributed localization scheme for mobile networks. In: *2015 IEEE 35th International Conference on Distributed Computing Systems (ICDCS)*, pp. 527–536. IEEE (2015)



A Fast Offline Database Construction Mechanism for Wi-Fi Fingerprint Based Localization Using Ultra-Wideband Technology

Huilin Jie, Hao Zhang, Kai Liu^(✉), Feiyu Jin, Chao Chen, and Chaocan Xiang

Department of Computer Science,
Chongqing University, Chongqing 400040, China
{jie0214,zhanghao1013,liukai0807,fyjin,cschaochen}@cqu.edu.cn,
xiang.chaocan@gmail.com

Abstract. With the ever-increasing demand on location-based services (LBS), fingerprint-based methods have attracted more and more attention in indoor localization. However, the considerable overhead of fingerprint is still a problem which hinders the practicability of such technology. Due to the prevalent of Wi-Fi access points (APs) and the high location accuracy of Ultra-Wideband (UWB), in this paper, we propose a hybrid system which utilizes UWB and Wi-Fi technologies to alleviate the offline overhead and improve the localization accuracy. Specifically, we employ UWB to determine the coordinate of each reference point (RP) instead of traditional manual measurement. Meanwhile, the Received Signal Strength Indicator (RSSI) of Wi-Fi is collected by a customized software installed in the mobile device. Then, a timestamp matching scheme is proposed to fuse these data coming from different devices and construct the offline fingerprint database. Besides, in order to better map the online data with offline database, an AP weight assignment scheme is proposed, which allocates APs with different weights based on the RSSI characteristic in each RP. We implement the system in real-world environment and the experimental results demonstrate the effectiveness of the proposed method.

Keywords: Indoor localization · Wi-Fi fingerprint · UWB technology

1 Introduction

The requirements of location based services (LBSs) have been growing rapidly along with the booming of intelligent terminals, which drives the development of indoor localization technology. In recent years the indoor localization has attracted more and more attention. In the outdoor space, especially, in the open area, the Global Positioning System (GPS) can well satisfy the demand of localization and navigation. However, the GPS based localization cannot be

applied in indoor environment due to severe signal attenuation. Therefore, many researchers pay attention to finding substitutive technologies for indoor localization, which contain common technologies such as Wi-Fi [1, 15, 20, 21], Bluetooth [2], RFID [3], infrared [4], and ultrasonic [5].

Compared with other technologies, Wi-Fi has been the hottest field in indoor localization due to the coverage area and the wide infrastructure deployment. Many indoor scenes are covered with Wi-Fi signal, which empowers Wi-Fi to conduct indoor location conveniently with the most commercial potential. However, the fluctuation of Wi-Fi RSSI signals due to a variety of environmental factors results in low localization precision. Therefore, many calibration works [16, 22], are required to assure the localization performance and maintain the robustness of the system. Ultra-wideband (UWB) [17] is another emerging indoor localization technology in recent years, which is a carrier-free communication technology. Due to high frequency band, the UWB possesses higher data transmission rate [18], which brings higher precision in line-of-sight (LOS) propagation. Although UWB technology can get more accurate positioning compared with Wi-Fi, it still suffers from non-line-of-sight (NLOS) scenes [19] (e.g., the block of wall, moving objects, etc.) and multipath effects because of the rectilinear propagation character. Besides, UWB technology has small coverage area, which needs more base stations for system deployment.

There are two mainstream location methods: range-based method and fingerprint-based method. The range-based method mainly relies on the measured distance between the receiver and the transmitter, then similar trilateration methods are employed to locate the target. However, this method is vulnerable to non-line-of-sight conditions and multipath effects caused by various interferences, which attenuates the location accuracy heavily. Clearly, such a technique is not suitable for the scenarios where with large-scale and high localization accuracy requirement. The fingerprint-based method contains two parts: offline training phase and online location phase. Compared with range-based methods, fingerprint is more robust. However, this method is time-consuming and labor-intensive in the offline phase, which requires a site-survey progress [23] to collect data at each preset reference point (RP) by the site surveyors. Besides, the maintenance cost of the whole system is high since the offline database needs to be updated frequently considering the ever changing of surrounding environments.

In this paper, considering UWB can achieve higher accuracy in indoor localization and the infrastructure of Wi-Fi is more widespread nowadays, the collaborative utilization of UWB and Wi-Fi is proposed. UWB is used to determine coordinate of each RP instead of manual estimation with the purpose of reducing offline database construction overhead. In detail, we first deploy UWB anchors and APs in the area. Then, we collect RSSI data and distance at each RP by Wi-Fi and UWB respectively. And the data (i.e., RSSI data and distances) are sent back to the server to build the offline database. Besides, considering the variation characteristic of Wi-Fi, we assign different APs with different weights for

better mapping online signal measurements at each RP. The main contributions of this work are outlined as follows.

- We propose a new hybrid system which synthetically exploits UWB and Wi-Fi technologies to conduct indoor localization. In detail, we utilize data fusion that combines coordinates collected by UWB with RSSI values provided by Wi-Fi to construct offline database. In the online localization phase, RSSI signal of Wi-Fi is collected to locate the target. The hybrid system reduces the overhead of site survey measurement.
- To construct offline database, we propose a timestamp matching scheme, which integrates the coordinate measured by UWB with RSSI value provided by Wi-Fi based on the time stamp of collected data. Besides, in order to better map online measurements with offline database, we propose an AP weight assignment scheme according to the character of RSSI data based on the observation that the variation degree of RSSIs at certain RP is changed along with the distance to the corresponding AP. On this basis, we assign different APs with different weights.
- We implement the system prototype on mobile device and carry out experimental validation in real-world environments. The results show the superiority of the proposed methods on reducing offline overhead and improving location accuracy.

The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 presents architecture of the proposed localization system. Section 4 introduces the algorithm. Comprehensive experiments and results are analyzed in Sect. 5. Finally, Sect. 6 gives the conclusion.

2 Related Work

Range-based method and fingerprint-based method are recently two mainstream methods in indoor localization. Range-based method mainly works out the distance between two sides of communications, including the time of arrival (TOA) method, angle of arrival (AOA) method etc. Hybrid TOA/RSS approaches have been researched in [6]. A modified MSE-based formula based on the hybrid TOA/RSS Cramér-Rao lower bound (CRLB) is implemented for more precise position estimation. Furthermore, they presented a novel hybrid TOA/RSS estimate approach based on a relaxation of the likelihood function, which shows outstanding performance among competitors. Yang et al. [7] improve the round-trip time (RTT) approach to obtain distance. After the transmitter sent out a burst of message, the receiver replies the response message to the transmitter at the preset time covering the time interval that the two sides had agreed on. However, obtaining measurement values such as TOA and AOA generally requires extra antennas support. In addition, non-line-of-sight (NLOS) conditions and multipath effects caused by various interferences have a terribly impact on measuring distance between transmitter and receiver, which heavily reduces localization accuracy.

The offline training phase and online location phase become a standard procedure for fingerprint-based method. Nevertheless, the offline database construction overhead is the most vital bottleneck. A site survey progress to collect data at RPs by the site surveyors is always required. Besides, the maintenance overhead of the whole system is high considering the ever changing of environment. LiFS [8] system introduces crowdsourcing methods to construct fingerprint database. Nevertheless, it needs accurate RSSI value collected at each RP to guarantee the localization precision. [9–11], use different methods reducing calibration measurements and handling heterogeneous devices. In [9], a region-partitioning mechanism is proposed that the positioning area is divided into small sub-areas by using dynamic linear boundaries between all pairs of APs. Each sub-area is associated with a unique AP-Sequence. However, it still needs some RP information by site survey as the reference for the first partition. Many researches, e.g. [10, 11], gradually focus on relative RSSI signals rather than absolute RSSI values. In [9], a binary RSSI gradient fingerprint database (Gmap) is constructed. The fingerprint is made up by the corresponding binary data instead of RSSI values, which reduces the overhead of maintaining fingerprint map. IncVoronoi [11] system is proposed which basic idea is relative RSSI signals received from two different APs can be mapped to relative distance. However, they just analysis RPs' RSSI characteristic but physical coordinates still need to be site surveyed.

Compared with the previous work, our system requires neither plenty of measurements of RPs by site surveyors nor a lot of computations in the training phase. Besides, we propose an AP weight assignment scheme to improve the indoor localization accuracy.

3 System Architecture

The Fig. 1 has shown the architecture of our hybrid system, which is based on fingerprint method consisting of offline phase and online phase. During the offline phase, distinguished from traditional offline database construction, UWB and Wi-Fi technologies are adopted to build offline fingerprint database. Actually, UWB has already reached centimeter-level location accuracy, however, users always need to carry a customized equipment to get their location, which limits its application scenario. Considering the high location precision of UWB technology, we utilize UWB to measure the coordinates of RPs instead of manual measurement in offline phase, which removes a lot of labor overheads. In this paper, as surveyor carried with an UWB transmitter module and a mobile device collects distances and RSSI respectively at each RP at the same time, the server continuously receives those two kinds data with respective timestamps. After that, these data stored in the server are used to build the offline database by a timestamp matching scheme. This scheme is devoted to combining the mixed data from different sources by timestamp. As a result of fusion, coordinates and RSSI which have the nearest timestamp comprise the fingerprint of each RP to construct offline database. Furthermore, an AP weight assignment scheme is

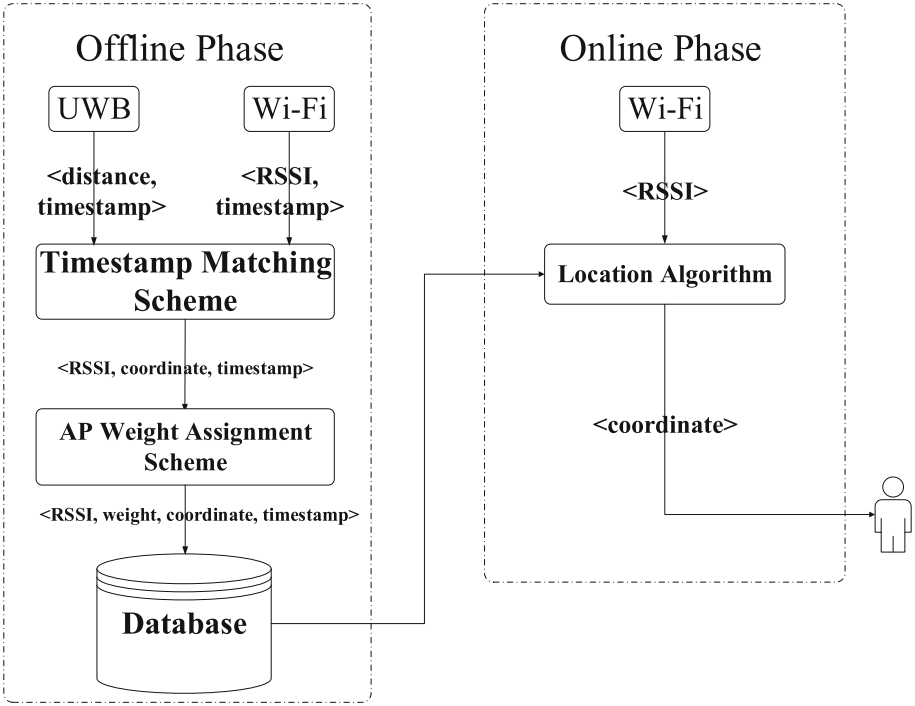


Fig. 1. System architecture

proposed based on the feature that variation degree of RSSI is changing with distance between the AP and target. This scheme aims at assigning each AP with different weights to better map the online RSSI values to offline fingerprints.

In the online phase, the collected RSSI data will be used to retrieve the neighboring data in the database and the coordinates of corresponding RPs are chosen to calculate the final coordinate of the target. On the basis of Euclidean distance, we calculate the distance between two sets of RSSI data with assigning different weights in each dimension. The detail will be elaborated in next section.

4 Indoor Hybrid Localization

4.1 The Construction of Offline Database

In the offline training phase, a site survey process is required to collect RSSI data and the corresponding coordinate at each RP, which is time-consuming and labor-intensive. In this paper, we employ UWB technology to measure the coordinate of each RP for removing traditional Wi-Fi calibration overhead, which reduces a lot of manual measurement work in a certain degree. The RSSI values provided by Wi-Fi and corresponding coordinates measured by UWB are integrated into offline fingerprints. Specifically, a UWB transmitter carried by the

site surveyor collects UWB signal, and a PAD is used to collect RSSI data. At least three UWB receivers acted as fixed anchors communicate with the UWB transmitter separately. The distances between each receiver and the transmitter are calculated by double-sided two-way ranging (DS-TWR) algorithm [14] in each UWB receiver module. This algorithm experiences more than once time of flight (TOF) progressing to address clock synchronization problem, which can obtain more accurate distance between two sides of communications. After gathered three ranging distances between the transmitter and three receivers, the UWB receiver which connects to the server transmits the distances to the server via serial port. Next, the server calculates the coordinates of sampling points by triangulation method. Due to the different sources of coordinate data and RSSI data, a timestamp matching scheme is required to fuse these data to construct the offline fingerprint database.

In detail, we set fixed time period to get reference points' coordinates provided by UWB. Considering fluctuation status and inaccuracy of RSSI values, the frequency of collecting RSSI is faster than obtaining coordinate at each RP. In consequence, the time period getting a coordinate from UWB would collect more RSSI values. In the purpose of making RSSI more accurate, we calculate the average value of the collected RSSI data at each RP. Then, the RSSI average value and one coordinate cooperate as one offline fingerprint, which constructs offline fingerprint database. Specifically, supposing there are total \mathbf{P} sets of RSSI values provided by Wi-Fi scanning and \mathbf{Q} coordinates values calculated by UWB modules, which are denoted by $RSSI = \{rssi_1, rssi_2, \dots, rssi_P\}$ and $C = \{c_1, c_2, \dots, c_Q\}$ respectively in the server. \mathbf{N} APs are installed in the environment, therefore, the i th $rssi_i$ can be represented by $rssi_i = \{rssi_{i1}, rssi_{i2}, \dots, rssi_{iN}, rssi_time_i\}$. The timestamp $rssi_time_i$ denotes the specific time at which i th $rssi_i$ is collected by mobile device. The j th coordinate is denoted by $c_j = \{x_j, y_j, coord_time_j\}$. Similarly, $coord_time_j$ represents the certain time at which the server calculated j th coordinate by using trilateration method with three ranging distances between receivers and the transmitter. We utilize the average of a fixed quantities of distances to calculate one coordinate. Therefore, the interval between the coordinates is a constant. Furthermore, considering the inaccuracy of RSSI values, the frequency of collecting RSSI values is higher than obtaining one coordinate, so one coordinate can get more corresponding RSSI values. Next, we search corresponding RSSI values for each coordinate by matching timestamp. In details, the time ranging Δ_j indicates the time between the $(j - 1)$ th timestamp $coord_time_{j-1}$ and the j th timestamp $coord_time_j$. After that, we search out the RSSI values from \mathbf{P} sets of RSSI values whose obtained timestamp are in the range of Δ_j . Supposing there are \mathbf{X} sets of RSSI values in the appointed range, which can be denoted by $RSSI = \{rssi_1, rssi_2, \dots, rssi_X\}$. Then we use the average of these \mathbf{X} sets of RSSI values to act as the j th final rssi values. Through this way, we can determine each coordinate with corresponding RSSI values. The final j th fingerprint can be represented as $f_j = \{\overline{rssi_{j1}}, \overline{rssi_{j2}}, \dots, \overline{rssi_{jN}}, x_j, y_j, coord_time_j\}$. As a

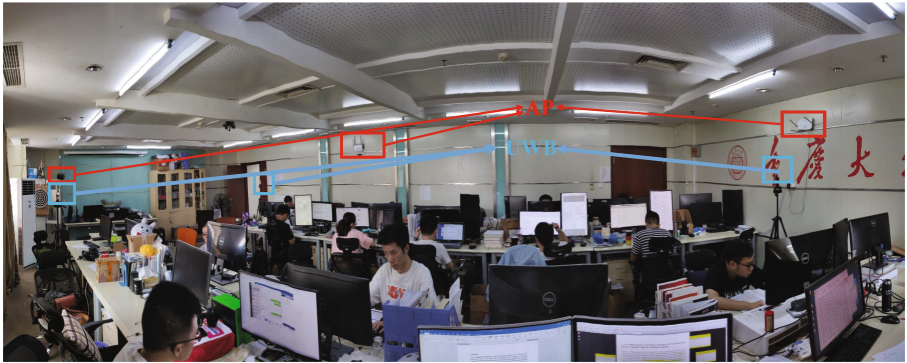
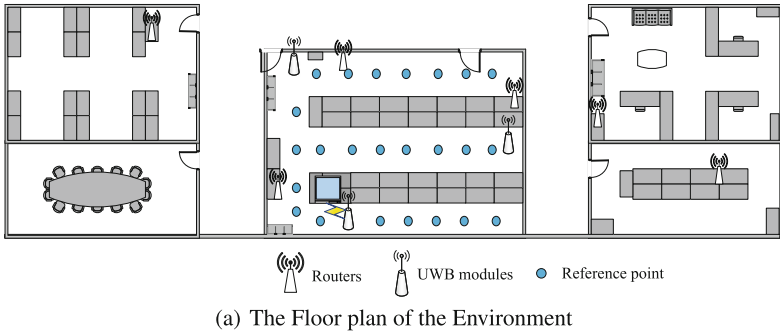


Fig. 2. The deployment of experiment

result, fingerprints which have been fused by our proposed timestamp matching scheme construct the offline database.

4.2 AP Weight Assignment Scheme

In online location phase, the measured RSSI data is used to retrieve the neighboring RPs in offline database by calculating Euclidean distance. This method treats equally each dimension of the RSSI data. Actually, the contribution of each dimension is different in distance calculation for indoor localization. In general, bigger the RSSI value is, closer the target is to the corresponding AP. However, the change of RSSI value does not obey this strictly due to the complex indoor interferences. In order to better describe the characteristic, we propose to use change frequency as the metric to evaluate the contribution of each AP.

In detail, supposing there are N APs and M RPs in the environment, which are denoted by $AP = \{ap_1, ap_2, \dots, ap_N\}$ and $RP = \{rp_1, rp_2, \dots, rp_M\}$, respectively. \overline{rssi}_{ij} represents the average RSSI value of i th AP and j th RP stored in offline database.

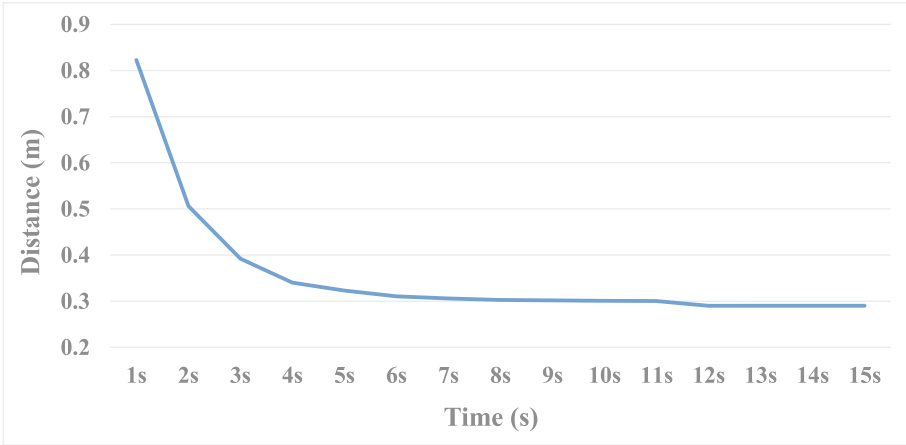


Fig. 3. Average UWB coordinate error varying with time

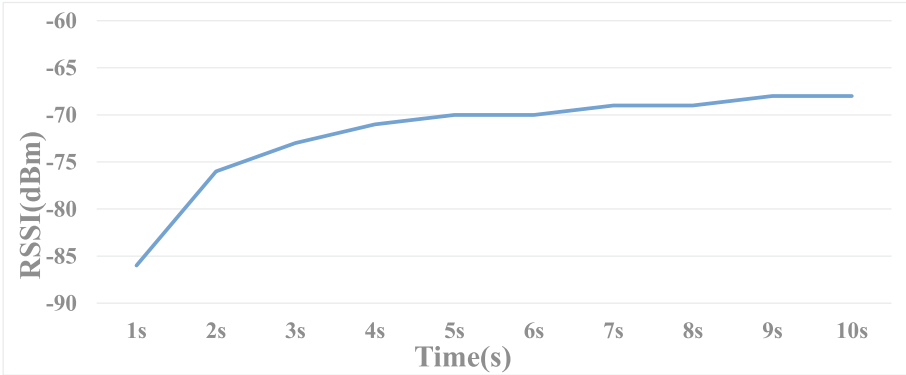


Fig. 4. Average RSSI value varying with time

We denote V_j as the change times in j th AP and initialize it as zero. If the current RSSI value is different from last status in this RP during offline sampling time, the value of V_j plus one.

Then, the weight w_j of j th AP are calculated as:

$$w_j = \frac{V_j}{\sum_{j=1}^N V_j} \tag{1}$$

Through this way, we can determine the weight of each AP in each RP. In the online location phase, we adopt KNN method to calculate the coordinate of the target. Specifically, when one online measurement $\mathbf{r} = (rssi_1, rssi_2, \dots, rssi_N)$ is received, the nearest K RPs can be filtered out by the Euclidean distance d_j , which is calculated by:

$$d_j = \sqrt{\sum_{i=1}^N (rssi_i - \overline{rssi_{ij}})^2 * w_i} \quad (2)$$

The selected RPs can be denoted by $RP' = \{rp'_1, rp'_2, \dots, rp'_K\}$. Then the final coordinate (\tilde{x}, \tilde{y}) of the target can be obtained by

$$\tilde{x} = \frac{1}{K} \sum_{i=1}^K x_i \quad (3)$$

$$\tilde{y} = \frac{1}{K} \sum_{i=1}^K y_i \quad (4)$$

where (x_i, y_i) is the coordinate of rp'_i .

5 Performance Evaluation

5.1 The Configuration of Experiment

We carry out the experiment in our laboratory, and the RPs are denoted with blue dots as shown in Fig. 2(a). There are near 20 students studying in the laboratory. The APs and UWB anchors are deployed in the surrounding with different icons. As shown in Fig. 2(b), the UWB modules are rested on tripods in the laboratory and the APs are fixed on the wall. We collect RSSI data at each RP lasting T seconds, and the corresponding coordinate is derived by the UWB at the same time.

5.2 Experimental Results

The Determination of Collection Time. In order to determine the collection time, we first explore the cumulative time that the average value of measured coordinates and RSSI signals can be stable. To this end, we collect these data in a fixed point with enough time and the cumulative average value of each second is computed. The results have been shown in Figs. 3 and 4.

From Fig. 3 we can see that along with time increasing, the distance errors between coordinates calculated by UWB and real coordinates gradually decline till they become stable. Specifically, at the seventh second, the distance stabilizes at 0.3m. Similarly, the RSSI maintains stability at -70 dBm with the cost of five seconds. Through this way, we choose the larger value (i.e., $T = 7$ s) of these two values for the purpose of guaranteeing the reliability of data.

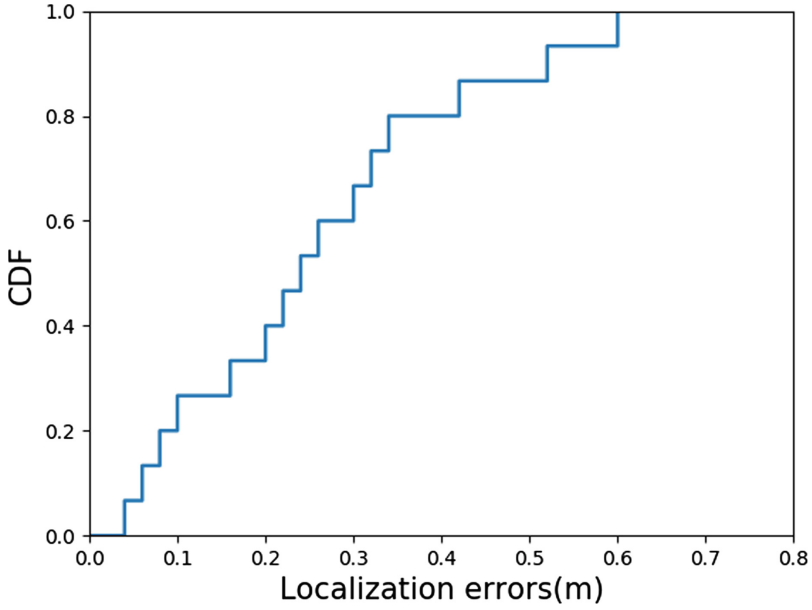


Fig. 5. CDF of UWB localization error

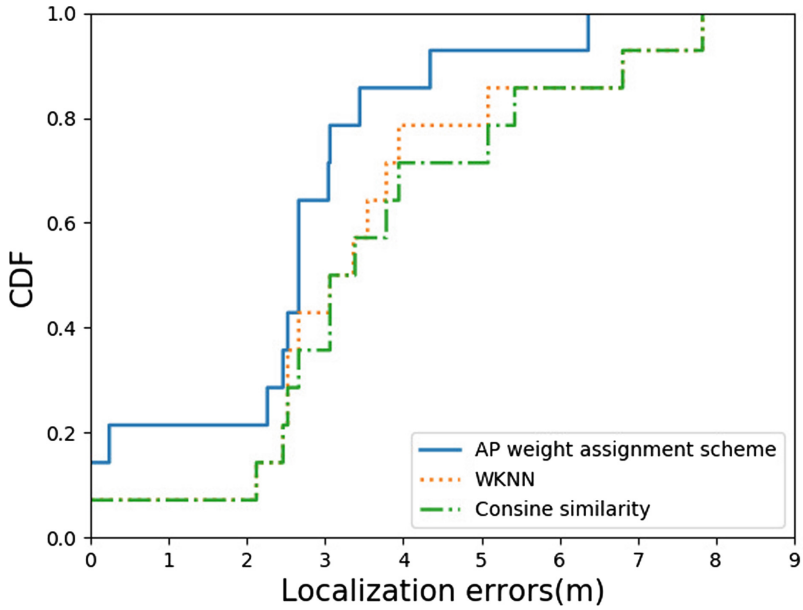


Fig. 6. CDF of localization error compared with other two algorithms

Accuracy Analysis of UWB. The location performance of UWB technology is evaluated, which affects directly the feasibility of the proposed method. We compare the UWB location results with the ground truth. At each RP, we sample the UWB signal with 7s and calculate the average value. In Fig. 5, we show the CDF of UWB localization error. This figure clearly shows that more than 85% of the test points have a localization error of 0.5m, which also validate the feasibility that replacing the manual measurement with the UWB localization results.

The Localization Result. We compare our proposed AP weight assignment scheme with other two existing algorithms which are WKNN algorithm [12] and weighted cosine similarity algorithm [13] in online location phase. The Fig. 6 shows the CDF of localization accuracy of these three algorithms. As shown, the proposed method achieves better performance than other two methods.

6 Conclusion

In this paper, we integrate UWB and Wi-Fi technologies to reduce offline overhead. Specifically, we utilize a customized software embedded on a mobile device to collect RSSI signal. In order to reduce the overhead in measuring the coordinate, we use UWB to derive the location information automatically instead of manual measurement. In order to fuse these data, we adopt a timestamp matching scheme to build the correspondence between RSSI and coordinate at each RP. After that, we propose an AP weight assignment scheme to improve the performance in indoor localization based on the RSSI signal changing feature. At last, the KNN is employed to estimate the final location of the target. We implement a prototype of the system and give a comprehensive testing in real-world environment. The experimental results show the effectiveness of the proposed method.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant No. 61872049; the Frontier Interdisciplinary Research Funds for the Central Universities (Project No. 2018CDQYJSJ0034); and the Venture & Innovation Support Program for Chongqing Overseas Returnees (Project No. cx2018016).

References

1. Shu, Y., Bo, C., Shen, G., Zhao, C., Li, L., Zhao, F.: Magicol: indoor localization using pervasive magnetic field and opportunistic WiFi sensing. *IEEE J. Sel. Areas Commun.* **33**(7), 1443–1457 (2015)
2. Zhuang, Y., Yang, J., Li, Y., Qi, L., El-Sheimy, N.: Smartphone-based indoor localization with bluetooth low energy beacons. *IEEE Sens.* **16**(5), 596 (2016)
3. Fang, Y., Cho, Y.K., Zhang, S., Perez, E.: Case study of BIM and cloud-enabled real-time RFID indoor localization for construction management applications. *J. Constr. Eng. Manag.* **142**(7), 05016003 (2016)

4. Wang, K., Nirmalathas, A., Lim, C., Alameh, K., Li, H., Skafidas, E.: Indoor infrared optical wireless localization system with background light power estimation capability. *Opt. Express* **25**(19), 22923–22931 (2017)
5. Yayan, U., Yucel, H.: A low cost ultrasonic based positioning system for the indoor navigation of mobile robots. *J. Intell. Rob. Syst.* **78**(3–4), 541–552 (2015)
6. Coluccia, A., Fascista, A.: On the hybrid TOA/RSS range estimation in wireless sensor networks. *IEEE Trans. Wirel. Commun.* **17**(1), 361–371 (2017)
7. Yang, C., Shao, H.-R.: WiFi-based indoor positioning. *IEEE Commun. Mag.* **53**(3), 150–157 (2015)
8. Yang, Z., Wu, C., Liu, Y.: Locating in fingerprint space: wireless indoor localization with little human intervention. In: *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, pp. 269–280. ACM, New York (2012)
9. Jun, J., et al.: Low-overhead wifi fingerprinting. *IEEE Trans. Mob. Comput.* **17**(3), 590–603 (2017)
10. Shu, Y., et al.: Gradient-based fingerprinting for indoor localization and tracking. *IEEE Trans. Ind. Electron.* **63**(4), 2424–2433 (2015)
11. Elbakly, R., Youssef, M.: A robust zero-calibration RF-based localization system for realistic environments. In: *2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pp. 1–9. IEEE, London (2016)
12. Machaj, J., Brida, P., Piché, R.: Rank based fingerprinting algorithm for indoor positioning. In: *2011 International Conference on Indoor Positioning and Indoor Navigation*, pp. 1–6. IEEE, Guimaraes (2011)
13. Han S., Zhao C., Meng W., Li, C.: Cosine similarity based fingerprinting algorithm in WLAN indoor positioning against device diversity. In: *2015 IEEE International Conference on Communications (ICC)*, pp. 2710–2714. IEEE, London (2015)
14. Silva, B., Pang, Z., Akerberg, J., Neander, J., Hancke, G.: Experimental study of UWB-based high precision localization for industrial applications. In: *2014 IEEE International Conference on Ultra-WideBand (ICUWB)*, pp. 280–285. IEEE, Paris (2014)
15. Martin, E., Vinyals, O., Friedland, G., Bajcsy, R.: Precise indoor localization using smart phones. In: *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 787–790. ACM, New York (2010)
16. Chai, E., Shin, K.G.: Low-overhead control channels in wireless networks. *IEEE Trans. Mob. Comput.* **14**(11), 2303–2315 (2015)
17. Toth, C.K., Jozkow, G., Koppányi, Z., Grejner-Brzezinska, D.: Positioning slow-moving platforms by UWB technology in GPS-challenged areas. *J. Surv. Eng.* **143**(4), 04017011 (2017)
18. Shi, G., Ming, Y.: Survey of indoor positioning systems based on ultra-wideband (UWB) technology. In: Zeng, Q.-A. (ed.) *Wireless Communications, Networking and Applications*. LNEE, vol. 348, pp. 1269–1278. Springer, New Delhi (2016). https://doi.org/10.1007/978-81-322-2580-5_115
19. Alarifi, A., et al.: Ultra wideband indoor positioning technologies: analysis and recent advances. *Sensors* **16**(5), 707 (2016)
20. Zhang, H., Liu, K., Jin, F., Feng, L., Lee, V., Ng, J.: A scalable indoor localization algorithm based on distance fitting and fingerprint mapping in Wi-Fi environments. *Neural Comput. Appl.* **2019**, 1–15 (2019)
21. Zhang, H., et al.: An Annulus Local Search Based Localization (ALSL) algorithm in indoor Wi-Fi environments. In: *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced*, pp. 887–892. IEEE, Guangzhou (2018)

22. Liu, K., et al.: Toward low-overhead fingerprint-based indoor localization via transfer learning: design, implementation, and evaluation. *IEEE Trans. Ind. Inform.* **14**(3), 898–908 (2017)
23. Jin, F., Liu, K., Zhang, H., Wu, W., Cao, J., Zhai, X.: A zero site-survey overhead indoor tracking system using particle filter. In: *ICC 2019–2019 IEEE International Conference on Communications (ICC)*, pp. 1–7. IEEE, Shanghai (2019)



A Moving Target Trajectory Tracking Method Based on CSI

Zhanjun Hao^{1,2(✉)}, Lihua Yan¹, and Xiaochao Dang^{1,2}

¹ College of Computer Science and Engineering, Northwest Normal University,
Lanzhou 730070, China

zhanjunhao@126.com

² Gansu Province Internet of Things Engineering Research Center,
Lanzhou 730070, China

Abstract. Aiming at the problems of high cost and low tracking performance of mobile target tracking, this paper proposes a CSI-based moving target trajectory tracking method. This method combines velocity estimation and hidden Markov model to achieve tracking of moving target trajectories. Firstly, the collected channel state information (CSI) in the offline phase, after preprocessing, is stored in the fingerprint database. Secondly, in the online stage, the model proposed in this paper is used for real-time matching, so as to realize real-time trajectory tracking of the target. Set up contrast experiments is carried out to verify the moving target trajectory tracking method proposed in this paper. The CSI-based moving target trajectory tracking method can track moving targets more accurately, has universality to different environments and targets, and has stability and robustness.

Keywords: Channel status information · Trajectory tracking · Velocity estimation · Hidden Markov model

1 Introduction

The rapid growth of location based services (LBS) has facilitated the rapid development of various positioning and tracking systems. Systems based on global positioning system (GPS) and cellular networks [1, 2] can provide high-precision positioning services in outdoor environments. However, due to the propagation barrier of GPS signals in indoor environments, These technologies cannot be directly used in indoor spaces. When describing human behavioral habits and tendencies, human walking trajectories are more complex than a single location. Therefore, a tracking based service (TBS) that presents services to users based on location and tracking is finer than LBS. Researcher use indoor positioning methods to obtain location to track the location of a range of users, but such methods typically require users to carry specialized devices such as mobile devices [3–5] and RFID [6]. Some researchers collect the moving target image according to the camera, and then use Kalman filter to track the moving target [7], but this method requires deploying many cameras, carrying equipment or deploying some cameras is inconvenient for the user. Wi-Fi-based systems, such as Wi-Fi [8], C²IL [9], Wi-Track [10] can track human rough motions and fine-grained

gestures, based on monitoring user actions can infer pedestrian movements track. However, the literature [8–10] must rely on an additional device, the tracking accuracy will be greatly reduced when the device changes.

In this paper, we proposed a CSI-based indoor tracking method. The system used two laptops with an Intel 5300 commercial wireless network card to collect CSI. The human body trajectory can be obtained by the hidden Markov tracking algorithm. The contributions of this paper are: Established a velocity estimation model, including velocity direction estimation, velocity magnitude estimation, and velocity correction. Established a hidden Markov tracking model to estimate the current state through the state of the previous phase. The moving target tracking method is verified in both the laboratory and the conference room, the tracking performance under different conditions, different packet transmission rates, different antenna heights, and different targets is compared.

The rest of this paper is structured as follows: The second part introduces related work. The moving target tracking method in this paper is introduced in detail in the third section. The fourth part includes the experimental environment and performance analysis of this paper. The last part summarizes the method of this paper.

2 Related Work

The successful application of GPS [11] in outdoor environments cannot be extended to indoor environments due to occlusion of buildings. Current indoor positioning methods fall into two categories: fingerprint-based and model-based. In the fire, firefighters need to be rescued in real time in the thick smoke. In hospitals, millions of valuable medical equipment need to be controlled; in the mall, consumers want to be able to navigate accurately to the shop or counter they want to go; old people or children, such people in need of protection, know in real time that their whereabouts can facilitate children or parents. So tracking-based location services are getting more and more attention.

In terms of device-based pedestrian trajectory tracking, researchers have different methods that require the target to carry sensors or specific tags, such as UWB tags [12], RFID tags [13], Bluetooth [14] devices, etc. Mobile devices or specific sensors to track human activity are inconvenient or even not possible in many cases. In view of the large fluctuation of the tracking trajectory caused by the instability of indoor Wi-Fi signals, Beihang Wang Fuwei et al. proposed the HMM-KFMC algorithm [15] to optimize the positioning tracking trajectory. Suraweera et al. proposed a passive tracking system with a decimeter level, Utilized an asynchronous self-positioning receiver to estimate TDOA to locate and track moving targets [16]. Literature [17] solved the problem of directional shadowing based on threshold-based fingerprint extraction. A trajectory tracking method based on map matching is proposed. In [18], a time-reversed indoor tracking method with centimeter-level accuracy is proposed. The time reversal technique is used to capture the difference in CSI, and then the mobile RF device is accurately located along its trajectory. Zhang et al. proposed an accurate indoor tracking system. Wi-Ball, which worked well in non-line-of-sight based on Wi-Fi signals [19]. In [20], an indoor position tracking system using a device-free passive (DFP) channel is proposed, which used fine-grained sub-channel measurement of MIMO-OFDM physical layer parameters to improve positioning and tracking accuracy.

Most of the current research on speed estimation by wireless signals is aimed at fast-moving targets, which are of great significance for safe driving [21–23]. In [24], a pedestrian speed estimation method based on OFDM system is proposed. Jiang et al. proposed an RSS-based speed estimation method under Wi-Fi combined with map matching to estimate the moving speed more accurately [17]. The most common used in velocity estimation is the Doppler algorithm, but the Doppler algorithm is not suitable for estimating human walking speed because the Doppler shift is negligible in a Wi-Fi environment with a maximum frequency of 5 GHz or 2.4 GHz. With the deepening of research, some researchers have integrated the Doppler shift into the MUSIC algorithm [25], which can estimate the pedestrian motion. Speed estimation based on dead reckoning has been applied to tracking systems. This paper combined velocity estimation with CSI signals and combines hidden Markov models to effectively track indoor moving targets.

3 Moving Target Trajectory Tracking Method

Due to the complex indoor environment, this article uses only two CSI-based commercial Wi-Fi devices: a data transmitter and a data receiver. This article used the following four steps to track the target trajectory. The first step is data collection, where the receiver receives Wi-Fi signals from the transmitter and records each CSI packet. The second step is to preprocess the CSI noise. The CSI data is easily interfered by external signals and generates a lot of noise. In this paper, the Principal Component Analysis and Fourier transform are used to process the CSI data. The third step is to track pedestrians based on the velocity estimation model, and establish a velocity estimation model and a velocity correction model to obtain a more appropriate speed. The fourth step is pedestrian’s trajectory can be tracked by establishing a Markov model. Figure 1 is a flow chart of moving target trajectory tracking.

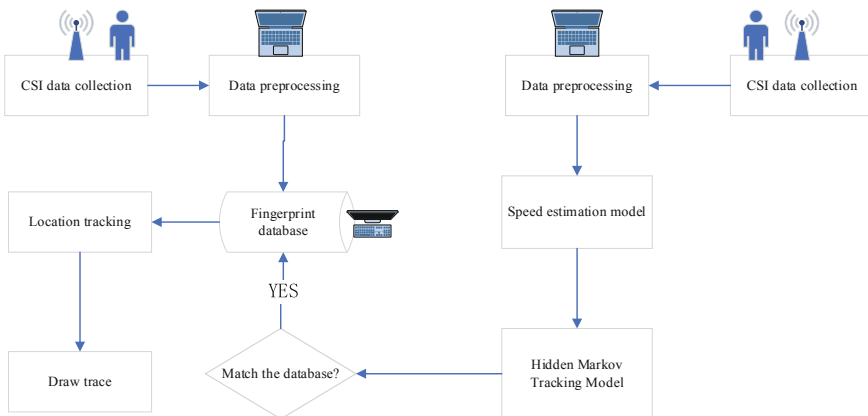


Fig. 1. Moving target trajectory tracking flowchart

3.1 Data Preprocessing

When there is no personnel movement in the indoor environment, there is a stable wireless signal propagation path, the CSI data is stable; when there is personnel movement in the room, the wireless signal is occluded by the target, so that the CSI data changes. Due to the complex indoor environment and various interferences, there is a lot of noise in the collected CSI data. In order to make the established fingerprint database match the real-time data more, the data needs to be denoised. This paper first used the wavelet transform, then extracted the CSI amplitude information by Fourier transform, and stored the processed data into the data fingerprint database. The upper left graph of Fig. 2 is the CSI raw amplitude data of 190 data packets, the upper right graph of Fig. 2 is the CSI amplitude data after the dimensionality reduction transposition, and the lower left graph of Fig. 2 is the amplitude map after wavelet processing, the lower right of Fig. 2 is grayscale. The data stored in the fingerprint database is the amplitude data after preprocessing.

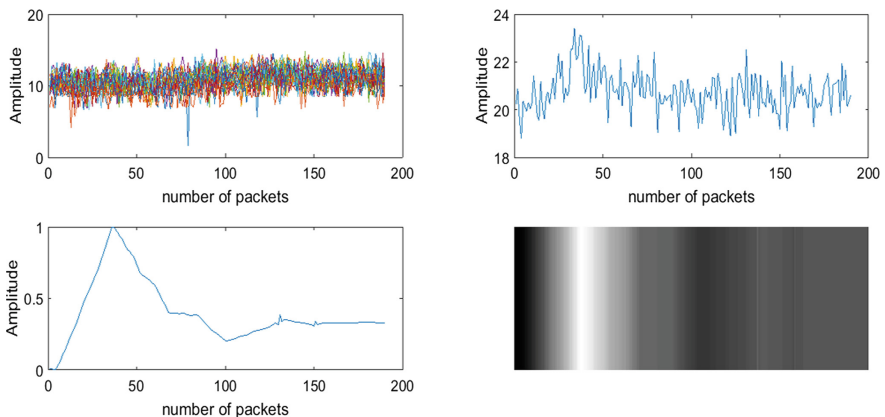


Fig. 2. Amplitude map

3.2 CSI-Based Speed Estimation

For the trajectory tracking of indoor moving targets, this paper proposed a CSI speed estimation model. Firstly, according to the CSI data collected by the target walking trajectory, the time of starting and ending the motion is determined and the motion direction is determined according to the speed direction estimation model. The velocity magnitude estimation model determines the velocity of motion and corrects it to obtain an accurate velocity estimate.

When there is no personnel movement in the room, the CSI data is stable due to the stable wireless signal propagation path, and when there is personnel movement, the CSI data changed greatly. Figure 3(a) and (b) are phase diagrams when the subject is stationary and walking along a straight line.

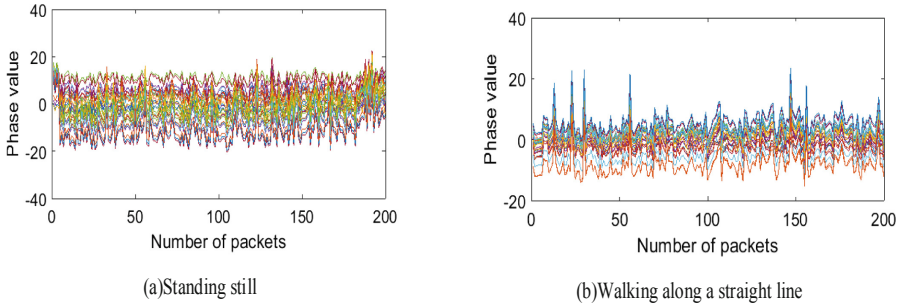


Fig. 3. Phase diagram

Figure 4(a) and (b) are subcarrier index map when the subject is stationary and walking along a straight line.

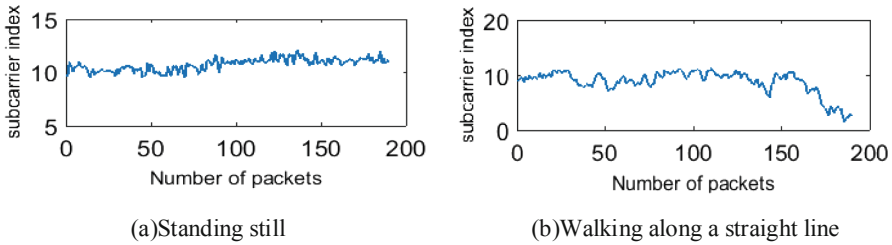


Fig. 4. Subcarrier index map

It can be seen from the figure that the amplitude at rest is basically stable, the amplitude changes greatly during exercise. Compared with Fig. 3(a) and (b), show that the change of the 50th package and the 150th package are obvious. The peak of the package set in this article is 100 packets per second, so it can be judged that the target starts to move at 0.25 s, the target continues to move at 0.5 s, and the target moved at a constant speed of 0.5 s to 1.5 s. Therefore, the node of the target motion can be judged from the phase map.

Estimation of Speed Direction. The method proposed in [28] is used to estimate the radial direction of human motion, which calculates the cross covariance between time-lag subcarrier fragments. We use the CSI segment when $T = 0.1$ s and calculate the time lag between subcarrier every 5 index intervals. Then, the radial direction of the moving target can be determined by calculating the distribution of the delay symbols. So we have the direction vector as $[dx, dy]^T$, dx and dy is the delay symbol accumulated in the 1 s time window. In this paper, the actual direction of human motion is given by direction synthesis. We have established a pair of mutually perpendicular transceiver links. dx and dy the radial directions of the two dimensions, respectively, (x_1, y_1) and (x_2, y_2) the midpoint coordinates of the transceiver link, respectively, (x_0, y_0) are the coordinate positions of the current target. So the direction angle of the speed can be expressed as

$$\theta_1 = \arctan\left(\frac{x_1 - x_0}{y_1 - y_0}\right) \quad (1)$$

$$\theta_2 = \arctan\left(\frac{x_2 - x_0}{y_2 - y_0}\right) \quad (2)$$

Speed Estimate. From the signal transmitting end to the signal receiving end, the CSI signal passes through multiple paths, including the LOS path and the path reflected by surrounding objects. Reference [26] establishes a model that associates CSI dynamics with path length change rate (PLCR). Because CSI can be represented by PLCR components. The specific steps are as follows:

Step1: The wavelet filter is used to obtain the amplitude information related to the moving target.

Step2: The first principal component is extracted from the filtered CSI data by Principal Component Analysis (PCA) and used as effective information for subsequent calculation. PCA not only gets the most relevant information for moving objects, it also reduces computational complexity.

Step3: Apply a short-term Fourier transform to the first principal component to obtain an amplitude phase and subcarrier index map.

Step4: Use the percentile method described in [19] to obtain reasonable PLCR data and estimate the speed of the moving target based on the PLCR data.

Speed Correction. The velocity magnitude and direction estimation model gives speed information to some extent, but it is necessary to further correct the radial velocity. Since the PLCR is affected by both the target speed and position, the calibration process needs to consider the location of the target.

Step1: Suppose the distance between the target and the midpoint of the transceiver is d , the actual radial velocity of v_0 . We observe the approximate logarithmic function between the estimated velocities v and d . Using the collected CSI to fit the logarithmic function f , we can get the following actual radial velocity:

$$v_0 = \frac{v}{f(d)} \quad (3)$$

Step2: We can solve the real speed according to the precise radial speed, When the target coordinate is (x_i, y_i) , the actual radial velocity is v_r , and the coordinates in the transceiver are (x_0, y_0) . The actual direction of the target motion has been obtained in the direction estimation and therefore β_1 known angle. β_2 can be obtained through a triangular relationship:

$$\beta_2 = -\arctan\left(\frac{x_i - x_0}{y_i - y_0}\right) \quad (4)$$

Obtain real speed

$$v = \frac{v_r}{\cos(\beta_1 + \beta_2)} \quad (5)$$

Step3: Find the average of several speeds to get the target speed. Combined with direction estimation and velocity estimation, the velocity estimation of moving targets can be solved in an indoor environment.

3.3 Hidden Markov Trajectory Tracking Algorithm

The Hidden Markov Models (HMM) is determined by the initial state vector π , the state transition probability matrix A , the observation probability matrix φ . π and A determine the sequence of states, φ determines the sequence of observations. The state transition probability matrix A and the initial state probability vector π determine the hidden Markov chain and generate an unobservable state sequence. The observation probability matrix φ determines how to generate observations from the state, and integrates with the state sequence to determine how to generate the probability sequence.

Three Elements of Hidden Markov Model. Initial state vector π , $\pi = \{\pi_1, \pi_2, \pi_3, \dots, \pi_n\}$, represents the probability that the model appears at each anchor point at the moment, since the initial state has equal probability at each point, so:

$$\pi_i = \frac{1}{n} \quad (6)$$

State transition probability matrix, $A = \{a_{ij}\}$, Where a_{ij} represents the probability that a pedestrian will move from i to j , which is $a_{ij} = \Pr(S_t = j | S_{t-1} = i)$, $1 \leq i, j \leq n$. Because the speed of pedestrians is generally not greater than 2 m/s, so only i and j are adjacent to the non-zero value, in other cases are zero. Assuming that the lattice point adjacent to i has an k_i , respectively j_1, j_2, \dots, j_{k_i} , then the non-zero transition probability is:

$$a_{ii} = a_{ij_1} = \dots = a_{ij_{k_i}} = \frac{1}{k_i + 1} \quad (7)$$

Through this method of establishing a matrix, the corresponding fingerprint library information is stored in the HMM.

Observation probability matrix ϕ , $\varphi = \{\varphi_1, \varphi_2, \dots, \varphi_n\}$ when $p_i - \hat{p} \neq 0$,

$$\varphi_i = \frac{1}{|p_i - \hat{p}|} \quad (8)$$

When $p_i - \hat{p} = 0$,

$$\varphi_i = C_g = \frac{2}{D_{RP}} \quad (9)$$

Where p_i is the coordinate of the i position, \hat{p} is the estimated coordinate of the resulting position, C_g is a sufficiently large constant value.

Markov Location Decoding Algorithm. In order to ensure the real-time nature of the trajectory positioning, each time an observation value is obtained, the local maximum probability at this time needs to be calculated, and the hidden state of the local maximum probability is taken as the estimated value at this moment.

The specific algorithm process are as follows: Calculate the local probability and hidden state of $t = 1$. Use initial probability when $t = 1$ and Corresponding observation state S_1 Calculation:

$$\delta(i, 1) = \pi_i \varphi_{is_i} \quad (10)$$

$$\psi_1 = \arg \max_{1 \leq i \leq N} \delta(i, 1) \quad (11)$$

Calculate the local probability $\delta(i, t)$ and hidden state of $t > 1$. Use the hidden state of the previous moment and the observation probability S_t of the observed state at this moment:

$$\delta(i, t) = q_{\psi_{t-1}i} \cdot a_{\psi_{t-1}i} \cdot \varphi_{is_t} \quad (12)$$

$$\psi_t = \arg \max_{1 \leq i \leq N} \delta(i, t) \quad (13)$$

Where $q_{\psi_{t-1}i}$ is the corrected probability of the transition probability, determine by:

$$q_{ij} = pr(v_t | S_t = j, S_{t-1} = i) = \frac{e^{\frac{1}{2}(v_t - \mu_{ij})^T \sum_v^{-1} (v_t - \mu_{ij})}}{2\pi \sqrt{|\sum_v|}} \quad (14)$$

The velocity component v_t is determined by the velocity estimation model.

4 Experimental Verification and Performance Analysis

4.1 Lab Environment

At present, CSI data is commonly used in Atheros 9380 and Intel 5300 models. The experimental equipment in this paper is two Lenovo desktops equipped with Intel 5300 network card. The operating system supports Ubuntu14.04.4 and installs Linux 802.11n CSI Tool. The CPU model is Intel Core i3-4150, one of which is a signal receiver and one is a signal transmitter. One of the experimental sites is the office area of 12 m × 8 m. Taking a 6 m × 6 m square sub-area in the office area of 12 m × 8 m as the experimental area. The sub-area is divided into 25 grids of 1.2 m × 1.2 m. The distance between the signal receiver and the signal transmitter is 5 m. The other experimental sites is conference room, the conference room is relatively empty, the layout is simple. Figure 5 shows the detailed deployment and plan of the conference room, and Fig. 6 shows the detailed deployment and plan of the more complex laboratory.

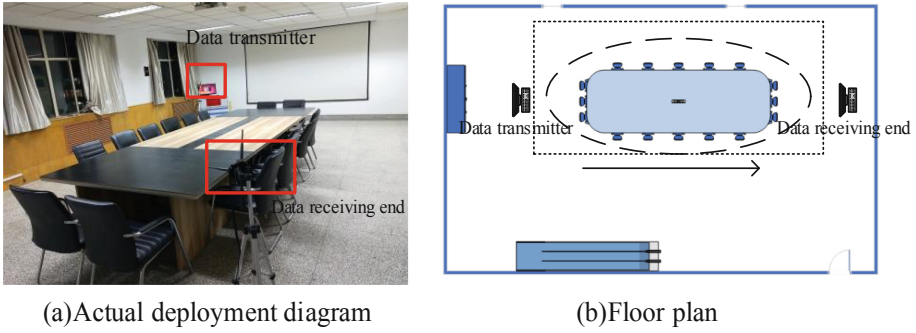


Fig. 5. Meeting room

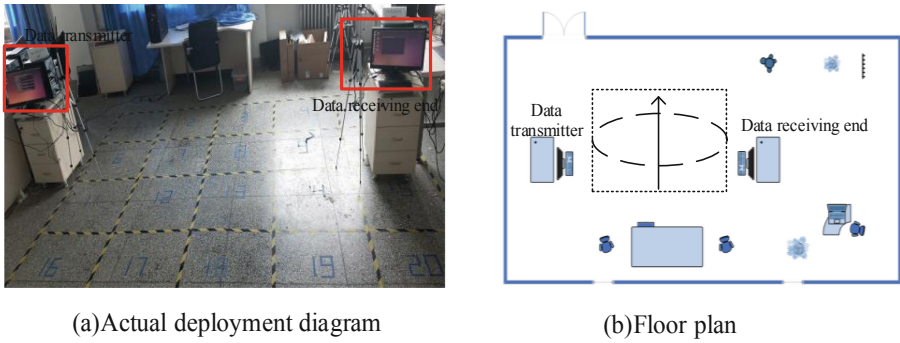


Fig. 6. Laboratory

4.2 Experimental Results

The pedestrian tracking method proposed in this paper was verified by experiments in the above two experimental environments. In the training phase, the aims is first allowed to collect CSI data at each reference point, each time collecting 500 packets, and collecting 100 times repeatedly, analyzing the data and pre-processing it and storing it in the fingerprint database; Walking according to a pre-defined path, and 100 times of data is collected for each path, and the processed data is analyzed and stored in the fingerprint database. In the online matching phase, all experimental equipment is kept in line with the training phase, enabling real-time tracking of pedestrians. The tracking results of the laboratory are shown in Fig. 7. Figure 7(a) is the tracking trajectory when the subject walks along a straight line, and Fig. 7(b) is the tracking trajectory when the subject walks along the rectangle.

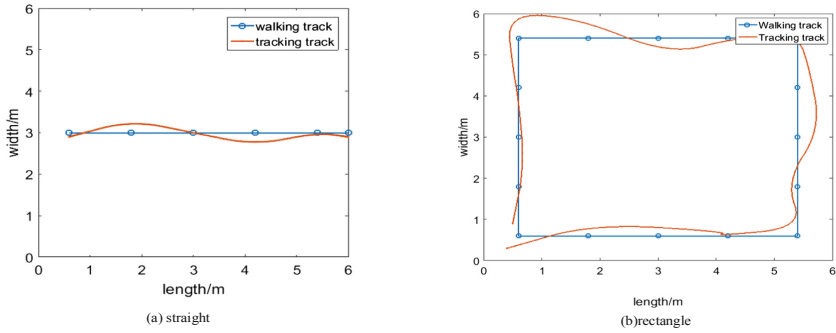


Fig. 7. Laboratory tracking results

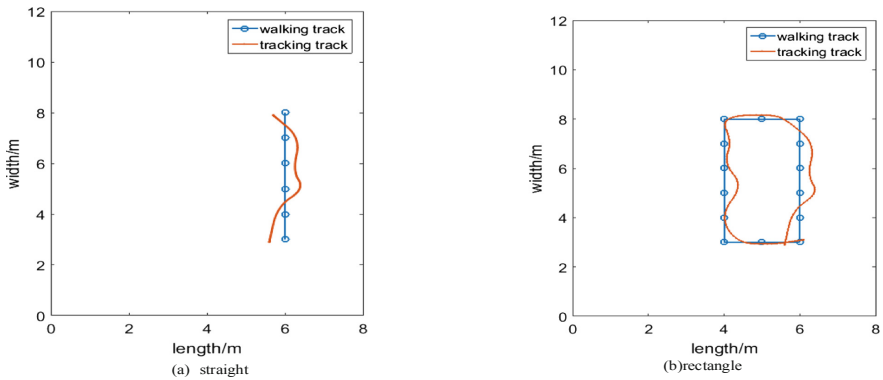


Fig. 8. Meeting room tracking results

The tracking result of the conference room is shown in Fig. 8. Figure 8(a) is the tracking trajectory when the subject walks along a straight line, and Fig. 8(b) is the tracking trajectory when the subject walks along the rectangle.

The tracking results shown in Figs. 7 and 8 show that in the open conference room, the multi-path effect is smaller and the tracking effect is more accurate. When the walking trajectory is a straight line, the tracking trajectory and the walking trajectory have a higher degree of coincidence, so that the simpler the walking trajectory, the better the tracking effect.

Figure 9 shows the cumulative distribution of tracking error in different experimental environments. The median error in a conference room with a relatively simple open environment is 0.93 m, and the median error is 0.97 m in a conference room with a large environmental complexity. The median error is only 0.4 m in different environments. The proposed method can track the pedestrian trajectory in different environments and has high robustness.

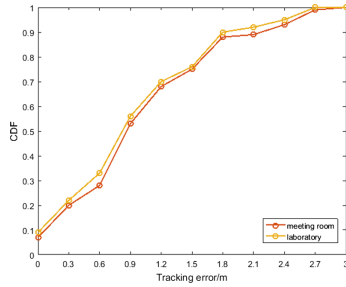


Fig. 9. Different environment comparison chart

4.3 Performance Analysis

In order to verify the stability and robustness of the proposed tracking method, a series of comparative experiments are designed to verify.

Impact of Packet Rate and Antenna Height. The packet sending rate has an impact on the collected CSI data. Different motion states correspond to different optimal packet sending rates. Therefore, in order to obtain the best tracking effect, the experiment verifies the effects of different packet sending rates, as shown in Fig. 10(a). Figure 10(b) shows the tracking accuracy of different antenna heights in both the laboratory and conference room scenarios.

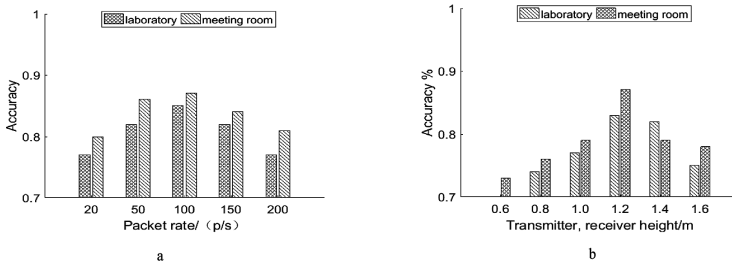


Fig. 10. Tracking accuracy map

Figure 10(a) compares the impact of the packet delivery rate on the accuracy of the tracking results in both the conference room and the laboratory. The accuracy of the laboratory is lower than that of the conference room. The accuracy is highest at a packet rate of 100 packets/second, the accuracy of the conference room is 88%, and the accuracy of the laboratory is 85%. Therefore, in order to achieve the best tracking effect, the subsequent comparative experiment fixed packet rate was 100 packets/sec. It can be seen from Fig. 10(b) that when the antenna height is 1.2 m, the tracking accuracy is the highest and the tracking accuracy is 87%. Therefore, the antenna heights of other comparative experiments in this paper are set to 1.2 m.

Impact of Different Speeds. In this paper, three different speeds are designed. The walking speed is less than 1 m/s for slow speed, the walking speed is between 1 m/s and 2 m/s for normal speed, and the walking speed is greater than 2 m/s for fast. Figure 11 shows the cumulative distribution of tracking error (CDF) at different speeds. Figure 11(a) is the laboratory speed error CDF chart, Fig. 11(b) is the conference room tracking error CDF chart.

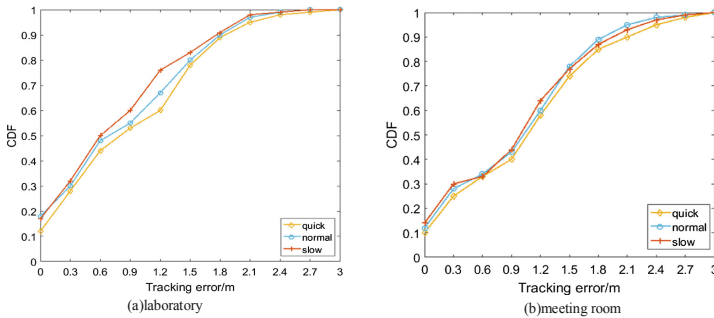


Fig. 11. CDF chart at different speeds

Figure 11(a) shows that the tracking error is not much different at different speeds. Fig. 11(b) shows the cumulative distribution of tracking error (CDF) at different speeds. As shown in Fig. 11(b), the slow, normal, and fast median tracking errors are 0.98 m, 1.02 m, and 1.05 m. It can be seen that the slower the speed, the better the tracking performance, but at different walking speeds, the tracking performance of this paper has achieved similar results. Comparing a and b, the curves are basically the same, so in different environments, the accuracy is 90% when the tracking accuracy is 2 m.

Impact of Walking Track. In order to verify the tracking performance under different trajectories, this paper designs four kinds of walking trajectories, which are straight lines, diagonal lines, circles and rectangles, and Fig. 14 shows the cumulative distribution of tracking error for different walking trajectories. Figure 14(a) is the laboratory error CDF chart, Fig. 11(b) is the conference room tracking error CDF chart.

Figure 12(a) shows that the tracking error of the laboratory is 75%, 78%, 81%, 85% within 2 m. As shown in Fig. 11(b), when the walking trajectories are straight lines, oblique lines, circles and rectangles, the probability of tracking error within 2 m is 89%, 85%, 83% and 79%. The simpler the trajectory, the higher the tracking performance and the more accurate the tracking results.

Impact of Different Targets. This paper selected four different targets for tracking experiments, namely 183 cm male, 178 cm male, 169 cm female, and 163 cm female. The experimental results are shown in Fig. 11(a). In real life, different pedestrians will have different walking trajectories

The four curves in Fig. 13 are basically coincident, and the different target tracking errors are basically the same, and the probability of tracking error within 2 m averages

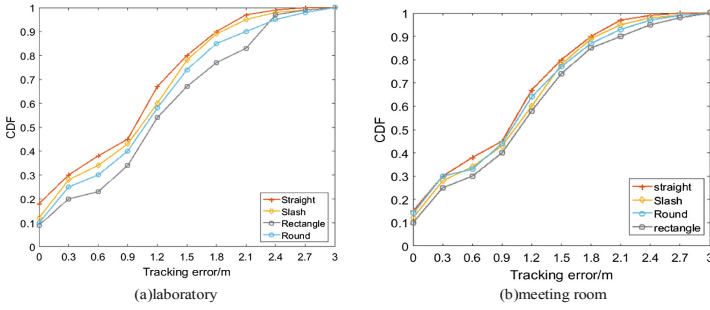


Fig. 12. CDF maps with different trajectories

87%. It can be seen that the tracking algorithm proposed in this paper is universal for different targets.

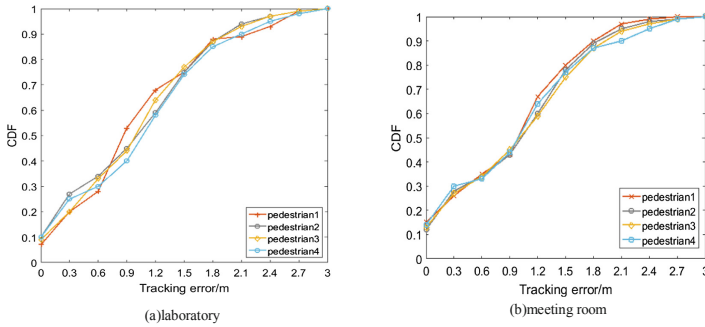


Fig. 13. CDF maps for different targets

Performance Comparison of Different Tracking Methods. In order to verify the moving target trajectory tracking method proposed in this paper, the advantages of the algorithm in tracking performance are reflected, and the method in this paper is compared with other target personnel tracking algorithms in Table 1.

Table 1. Tracking method performance comparison

Tracking method	Calculating time	Position mean square error	Tracking accuracy
SE-HMM	19.37	12.56	83.2%
Doppler-music	19.85	11.78	78.5%
K-Means	22.57	9.55	74.5%
KNN	24.89	8.43	64.3%

Table 1 shows the performance parameters of the moving target trajectory tracking method and other target tracking methods mentioned in this paper. It can be seen from

the table that the moving target trajectory tracking method proposed in this paper improves the tracking accuracy and reduces the calculation time and communication overhead. The tracking method of this paper is 6.7%, 8.7% and 11.9% higher than the Doppler-music algorithm, the traditional KNN and K-means algorithms, respectively. In general, the tracking performance has been improved, and the trajectory tracking of moving targets can be better realized.

5 Conclusion

Aiming at the problems of high tracking cost and low tracking performance of current mobile targets, this paper proposed a CSI-based moving target trajectory tracking method. The CSI data collected in real time and established speed estimation model and hidden Markov tracking model are used to track the moving targets in real time. Repeated experiments prove that the tracking method proposed in this paper can track the moving target, and the tracking median error is within 1 m, which effectively improves the tracking accuracy. Experiments with different environments, different targets and different speeds prove that the proposed method is universal and robust.

References

1. Li, Y., Zhu, X., Jiang, Y., Huang, Y., et al.: Energy-efficient positioning for cellular networks with unknown path loss exponent. In: 2015 IEEE International Conference on Consumer Electronics - Taiwan, Taipei, pp. 502–503 (2015)
2. Bulusu, N., Heidemann, J., Estrin, D.: GPS-less low-cost outdoor localization for very small devices. *IEEE Pers. Commun.* **7**(5), 28–34 (2000)
3. Chapre, Y., Ignjatovic, A., Seneviratne, A., et al.: CSI-MIMO: indoor Wi-Fi fingerprinting system. In: 39th Annual IEEE Conference on Local Computer Networks. IEEE (2014)
4. Shi, X., Ji, Z.: A radio frequency identification indoor tracking algorithm based on improved particle filter. *Comput. Eng.* **41**(11), 308–313 (2015)
5. Wu, K., Xiao, J., Yi, Y., et al.: FILA: fine-grained indoor localization. In: Proceedings - IEEE INFOCOM, pp. 2210–2218 (2012)
6. Shan, G., Feng, Y.: Video-assisted passive RFID indoor tracking technology. *Softw. Eng.* **19**(7), 18–21 (2016)
7. Qiao, K., Guo, C., Shi, J.: Research on moving human body tracking algorithm based on Kalman filter. *Comput. Digit. Eng.* **40**(1), 1–3 (2012)
8. Huang, G., Hu, Y., Cai, H., et al.: Wi-Fi fingerprint based indoor positioning method for smartphones [J/OL]. *Acta Automatica Sinica* 1–12, 30 April 2019. <https://doi.org/10.16383/j.aas.2018.c170189>
9. Jiang, Z.P., Xi, W., Li, X., et al.: Communicating is crowdsourcing: Wi-Fi indoor localization with CSI-based speed estimation. *J. Comput. Sci. Technol.* **29**(4), 589–604 (2013)
10. Adib, F., Kabelac, Z., Katabi, D., Miller, R.C.: 3D tracking via body radio reflections. In: USENIX NSDI, vol. 14 (2014)
11. Liu, J., Priyantha, B., Hart, T., et al.: Energy efficient GPS sensing with cloud offloading. In: Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems, pp. 85–98, November 2012

12. Zhu, C.: Comparative analysis of UWB indoor location tracking algorithm. Academic Communication Center of China Satellite Navigation System Management Office. In: Proceedings of the 8th China Satellite Navigation Academic Annual Conference - S02 Navigation and Location Service. Academic Exchange Center of China Satellite Navigation System Management Office: Organizing Committee of China Satellite Navigation Academic Annual Conference, p. 5 (2017)
13. Ni, L.M., Liu, Y., Lau, Y.C., et al.: Indoor location sensing using active RFID. *Wirel. Netw.* **10**(6), 701–710 (2004)
14. Feldmann, S., Kyamakya, K., Zapater, A., et al.: An indoor bluetooth-based positioning system: concept, implementation and experimental evaluation, pp. 109–113 (2003)
15. Wang, F., Huang, Z.: Research on tracking of moving targets in indoor positioning. *J. Naut. Navig.* **4**(01), 33–37 (2016)
16. Suraweera, N., Li, S., Johnson, M., et al.: A passive tracking system with decimeter-level accuracy using IEEE 802.11 signals. *Military Communications* (2018)
17. Jiang, Z.P., Xi, W., Li, X., et al.: Communicating is crowdsourcing: Wi-Fi indoor localization with CSI-based speed estimation
18. Chen, C., Han, Y., Chen, Y., et al.: Time-reversal indoor positioning with centimeter accuracy using multi-antenna WiFi. In: *Signal & Information Processing. IEEE* (2017)
19. Zhang, F., Chen, C., Wang, B., et al.: WiBall: a time-reversal focusing ball method for indoor tracking. *IEEE Internet Things J.* **PP**(99) (2017)
20. Shi, S., Sigg, S., Chen, L., et al.: Accurate location tracking from CSI-based passive device-free probabilistic fingerprinting. *IEEE Trans. Veh. Technol.* **PP**(99), 1 (2018)
21. Wei, X., Wang, X., Jin, J.: A method for estimating ship's Azimuth velocity based on local center frequency for SAR images. *J. Electron. Inf. Technol.* **40**(09), 2242–2249 (2018)
22. Wang, W., Wang, P., Su, W., et al.: A high speed target parameter estimation algorithm based on frequency domain super resolution. *J. Electron. Inf. Technol.* **38**(12), 3034–3041 (2016)
23. Lu, F., Chen, S., Liu, C., et al.: Estimation of vehicle vibration velocity based on Kalman Filter. *J. Vibr. Shock* **33**(13), 111–116 (2014)
24. Pricope, B., Haas, H.: Experimental validation of a new pedestrian speed estimator for OFDM systems in indoor environments. In: Proceedings of the 54th IEEE Global Communications Conference, December 2011
25. Hao, Z., Li, B., Dang, X.: A person trajectory tracking method based on channel state information [J/OL]. *Comput. Appl. Res.* **2019**(10), 1–3, 9 January 2019. <http://kns.cnki.net/kcms/detail/51.1196.TP.20180913.1708.002.html>
26. Qian, K., Wu, C., Yang, Z., et al.: Widar: decimeter-level passive tracking via velocity monitoring with commodity wi-fi. In: Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing, p. 6. ACM (2017)
27. Dorp, P.V., Groen, F.C.A.: Feature-based human motion parameter estimation with radar. *IET Radar Sonar Navig.* **2**(2), 135–145 (2008)
28. Wu, D., Zhang, D., Xu, C., et al.: WiDir: walking direction estimation using wireless signals. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 351–362. ACM (2016)



A CSI-Based Indoor Intrusion Detection and Localization Method

Xiaochao Dang^{1,2}, Caixia Li¹, Zhanjun Hao^{1,2}(✉), and Yuan Cao¹

¹ College of Computer Science and Engineering, Northwest Normal University,
Lanzhou 730070, China

zhanjunhao@126.com

² Gansu Province Internet of Things Engineering Research Center,
Lanzhou 730070, China

Abstract. In this paper, we propose a method for indoor intrusion detection and localization that makes use of channel state information (CSI), which consists of an offline phase and an online phase. In the former, we collect CSI in different scenarios, and at different times, for more comprehensive characterization of signal propagation. To reduce the redundancy and dimensionality of CSI data, we employ the principal component analysis algorithm to extract the main features of CSI, and build the fingerprint database for localization. In the online phase, we first apply the earth mover's distance algorithm to detect the presence of the person in the test area. Following this, we determine the approximate location of the target according to the change of CSI measurements, and compare this to the fingerprint database, to select reference points to build the sub-fingerprint database. Finally, we evaluate the actual position of this target using the improved k-Nearest Neighbor algorithm.

Keywords: Channel state information · Fingerprint database · Intrusion detection · Indoor localization

1 Introduction

In recent years, the demand for location-based services (LBS) has driven the rapid development of location-detection technology [1]. While global positioning systems (GPS) provide meter-level detection accuracy outdoors, the influence of non-line-of-sight (NLOS) propagation makes it difficult to achieve high-precision positioning with GPS in indoor environments. Therefore, a large number of indoor localization methods have been proposed, based on technologies such as Bluetooth [2], infrared sensor [3], smartphone sensors [4], radio frequency identification (RFID) [5], Wi-Fi [6], visible light communications [7], and Ultra-wideband [8]. Although most intrusion detection and localization methods require additional electronic devices and specialized hardware [9], in many cases, solutions where the use of these supplementary materials in detection is avoided, that is, passive detection and indoor localization techniques [10, 11], are necessary. The increasing prevalence of Wi-Fi has meant that passive detection and localization techniques based on this technology, which would reduce costs as well as improve their accuracy, are being studied intensively [12, 13].

The received signal strength (RSS) of Wi-Fi module is an easily accessible indicator that can be used for device free intrusion detection and localization [14, 15] because it varies with propagation distance. However, this metric is affected by multipath effects, which are more typical in complex indoor environments. To achieve multipath propagation, we can obtain channel state information (CSI) from some network interface cards (NIC) with orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) techniques. With advanced NICs, such as Intel Wireless Link 5300 (IWL5300) [16], CSI can be obtained trivially, and as such, it can be used for device-free detection in a similar manner to RSS. In addition, this indicator provides sub-carrier level channel measurement, which is helpful for improving the accuracy of intrusion detection and indoor fingerprint localization [17, 18], because physical layer CSI includes details such as the amplitude and phase of each sub-carrier in a channel, changes of the Wi-Fi signal occurring between the transmitter and the receiver are described better. As CSI measurements are subject to interference in more complex environments, some processing is required, using techniques such as principal component analysis (PCA) [19], Kalman filtering [20], and density-based spatial clustering of applications with noise [17], to establish fingerprint database (i.e., a record of characteristic CSI measurements). To complete the localization process, support vector machine (SVM) [21], Naive Bayes Classification [22, 23] and deep learning [13], Convolutional Neural Network [24, 25] algorithms can be applied to online match of CSI information to the fingerprint database.

In this paper, we propose an intrusion detection and localization method based on CSI (named DLFi). In the training phase, we first collect CSI data from a range of reference points, in two different scenarios. To improve the precision of detection and localization in complex environments, we employ the PCA algorithm for noise and dimensionality reduction, to extract the main features of the CSI and build fingerprint database. In the test phase, intrusion detection is achieved by comparing the CSI collected to information stored in the fingerprint database, using the earth mover's distance (EMD) algorithm. Following successful detection, we apply a localization algorithm to evaluate the position of the intruder. In this localization phase, the approximate location of the target is determined according to the detection result, and reference points are selected to build a sub-fingerprint database. Finally, the accurate position is evaluated using the improved k-Nearest Neighbor (kNN) algorithm. In the experiment validation section, we verify the performance of DLFi in two typical indoor environments: a spacious meeting room, and a crowded laboratory. Finally we compare DLFi with Nuzzer [26], a passive RSS-based technique, and Pilot [27], a passive CSI-based method. The main contributions of the paper can be summarized as follows:

- (1) We collect data packets from the IWL5300 NIC using the modified device driver, to obtain CSI. We extract only the signal amplitudes from the CSI data, for using as fingerprints.
- (2) We employ the PCA algorithm in the offline phase, for noise and data dimension reduction. In addition to improve the accuracy of localization, the duration of estimation is also reduced.

- (3) In the online phase, intrusion detection is judged using the EMD algorithm, the results of which are used to estimate an approximate area for further localization. We use the improved kNN algorithm to determine the exact position of a detected intruder.
- (4) We verify DLFi in two typical environments, and the results show that the average intrusion detection rate is more than 90% in both environments considered (a spacious meeting room and a crowded laboratory). The localization accuracy of DLFi is an improvement compared with those reported with the Nuzzer and Pilot methods.

2 Intrusion Detection and Localization

Intrusion detection refers to the passive detection of intruder in a specific area, usually without the permission of the intruder. But many additional devices need to be deployed to determine whether there is an intrusion through comparing information changes extracted from various devices, and then locate the intruder. However, these devices are usually expensive, and difficult to achieve widespread. Wi-Fi signals are ubiquitous in indoor environments. As long as there is Wi-Fi coverage, it is suitable for intrusion detection using Wi-Fi, which can achieve better result and save cost, while not requiring additional equipment support. In this paper, we use the Wi-Fi signal for intrusion detection and localization. In addition, in order to improve the accuracy of positioning, we narrow the target area based on the intrusion detection result, and then accurately locate the intruder in a small range. All of the above are main motivations for this work, and we will detail our work in the next step.

The CSI refers to the characteristic of a channel in a specific frequency band, and describes how the signal travels from the transmitter to the receiver. During the training phase, we collect measurements from a range of reference points to build fingerprint databases consisting of CSI amplitudes. To obtain this, the NIC collects CSI from the Wi-Fi signal, which has been modulated onto each sub-channel using the OFDM technique, to enable multipath transmission. The signal received following this can be expressed as:

$$Y = H \cdot X + N \quad (1)$$

Where Y and X are the respective signal vectors at the receiver and the transmitter, and H and N refer to the CSI matrix and Gaussian white noise, respectively. The amplitudes are extracted from the CSI data set, which are used as fingerprint features. At the offline phase, CSI data are preprocessed for fingerprint database construction. To improve the performance of DLFi, CSI measurements obtained from the NIC are processed using the PCA algorithm, which extracts the main signal features of these reference points. We selected this algorithm because, in contrast to other filtering techniques, it can reduce the dimensions of measurement data, as well as the magnitude of noise within these data (the noise is often obtained during the data collection process). This reduction in data dimensions is required as the complexity of a

characteristic system increases if statistical methods are applied to analysis of multi-variate data with a large number of variables. As each pair of transmit and receive antenna has one channel, and each channel of the IWL5300 network card contains 30 subcarriers, the dimensionality of the raw CSI data is relatively high. The PCA algorithm is able to identify correlations between different variables (which often indicate a shared characteristic), and remove redundant variables from dataset, reducing the overall complexity of the statistically derived relationship. Reducing the complexity of the CSI dataset also makes processing quicker, as calculation is more efficient.

Test CSI measurements are collected in the online phase, in which the EMD algorithm is used to detect intruders. The results of this detection provide a rough estimate of the location of the intruder. This information is subsequently compared with the data stored in the fingerprint database, to evaluate their similarity. Finally, we calculate weights reflecting this similarity, and estimate the accurate position of the intruder using a weighted kNN algorithm based on the Gaussian kernel function. The architecture of DLFi is shown in Fig. 1.

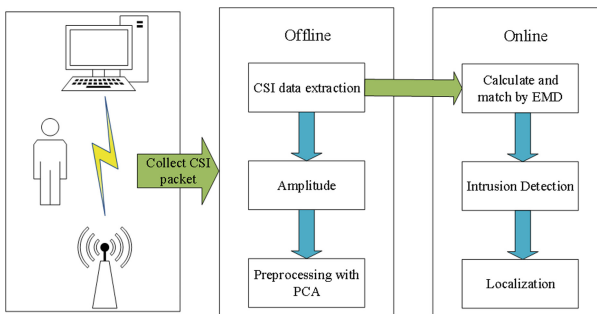


Fig. 1. Architecture of the DLFi method.

2.1 Data Process and Fingerprint Library Construction

As mentioned above, CSI data usually contain noise and the dimensionality is high, so the PCA algorithm is used to data process to filter out redundant data, which due to environmental changes and other factors, and extract main features. We collect 50 packets CSI data to test, the data comparison before and after processing is illustrated in Fig. 2. CSI amplitudes of 50 data packets are shown in Fig. 2(a), while Fig. 2(b) illustrates the main features extracted from 50 data packets, we can find that the CSI major features are extracted from multiple sets of data, which will be use as fingerprint to store in the data library.

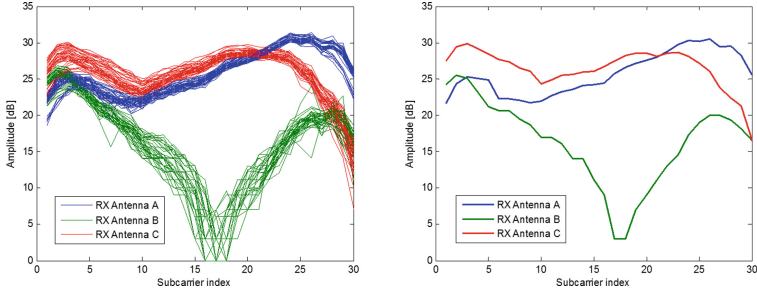


Fig. 2. CSI data comparison before and after processing.

During the offline phase, the fingerprint library is built after data process. In addition, to avoid interference from other factors, the indoor environment remains static, and only one tester stands at each reference point during data collection. The fingerprint database F is expressed as $F = \{f_1, f_2, \dots, f_i\}$, where i refers to the index of the reference point, and the coordinates of the reference point are given as (x_i, y_i) . In this equation, f_i represents the signal characteristic of each reference point, written as $f_i = \{(x_i, y_i), csi_i\}$. Especially, f_0 represents the fingerprint data while the room is vacant.

2.2 Intrusion Detection

CSI measurements are sensitive to the appearance of people in a room, as demonstrated by the different patterns illustrated in Fig. 3. Figure 3(a) shows the CSI amplitude of each subcarrier when the room is vacant, while Fig. 3(b) and (c) depict the patterns observed with an intruder in location 1 and location 2, respectively. There are obvious changes in the appearance of the CSI amplitudes, demonstrating that intrusion detection based on this metric is feasible.

In the intrusion detection stage, the EMD algorithm is employed to calculate the similarity between CSI measurements collected in online and offline phases of operation. In the offline phase, the similarity between the data collected from two receivers is subsequently calculated, and denoted as EMD_0 . In the online phase, data sets are taken from two receivers and denoted as CSI_1 and CSI_2 , respectively. We calculate the similarity between CSI_1 and CSI_2 , using the following equation:

$$EMD(CSI_1, CSI_2) = \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} d_{ij}}{\sum_{i=1}^m \sum_{j=1}^n f_{ij}} \quad (2)$$

Where $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = \min \left\{ \sum_{i=1}^m CSI_1, \sum_{j=1}^n CSI_2 \right\}$, And then calculate $EMD(CSI_1,$

$CSI_2)$ using the same method. Finally, determine $\min\{EMD_1, EMD_2\}$, and compare this with EMD_0 . If $\min\{EMD_1, EMD_2\} < EMD_0$, there is someone in the room, otherwise, the room is vacant.

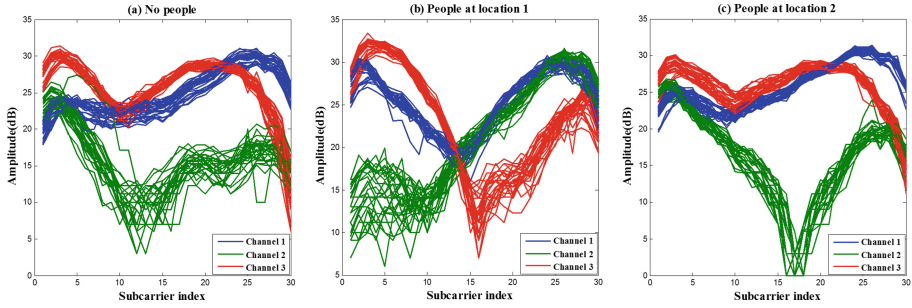


Fig. 3. CSI measurements exhibit different patterns depending on the absence and presence of people in the area of interest. (a) Area of interest is vacant; (b) People are present in location 1; (c) People are present in location 2.

2.3 Localization

The exact position of the intruder is evaluated in the localization stage. To improve the accuracy of localization, we use one access point (AP) and two monitor points (MPs) denoted as MP_1 and MP_2 . In this paper, we choose to use two MPs. In fact, we also considered using more MPs, because the more MPs, the more accurate the positioning results, but the more data that need to be processed. In the DLFi method proposed in this paper, the test area is refined by the intrusion detection result, and then the position is evaluated accurately. Through many experiments and verification analysis, we choose to use two MPs in order to save time and improve accuracy.

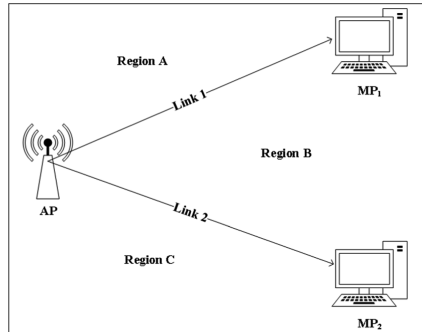


Fig. 4. Illustration of the target region, and corresponding sub-regions.

Figure 4 illustrates the geometry of the target region, which determines how localization is achieved. The target region is divided into three parts, marked as Region A, B, and C. While data collected from MP_1 and MP_2 when the room is vacant are very similar, the variation in the CSI amplitudes at both MPs differs if an intruder is present. This fact forms the basis of our localization technique. After indoor intrusion has been confirmed, we compare the CSI amplitudes collected in the online phase with

the data stored in the fingerprint database. Obvious changes in only the CSI amplitudes received from MP_1 illustrate that propagation in Link 1 has been affected, and the target is located in Region A. Conversely, obvious changes in only the CSI amplitudes received from MP_2 illustrates that Link 2 has been affected, and the target is located in Region C. Finally, differences between the CSI data collected from MP_1 and MP_2 and the data stored in the fingerprint database demonstrate that propagation in both Link 1 and Link 2 have been affected, placing the target in Region B.

Once the approximate location of a target has been determined, we select reference points to construct a sub-fingerprint database, with which we determine the precise position of the target. Diminishing the target region in this way not only reduces the duration of estimation, it also improves localization accuracy. In this stage, we use the Gaussian kernel function to calculate weights reflecting the similarity between test and reference data. The calculation method is as follows:

$$w_k = \frac{\phi(csi_t, csi_i)}{\sum_{i=1}^p \phi(csi_t, csi_i)} \quad (3)$$

$$\phi(csi_t, csi_i) = \exp\left(-\frac{\|csi_t - csi_i\|^2}{2\sigma^2}\right) \quad (4)$$

where w_k represents the weight, p is the number of reference points in the similarity set $Q(p)$, csi_t refers to CSI data collected in the test phase and csi_i represents stored in the fingerprint database, and σ is the parameter of the Gaussian kernel function. Finally, we determine the precise position of the target using the previously estimated weights and the weighted kNN algorithm. This position is calculated as follows:

$$\hat{L} = \sum_{k \in Q(p)} w_k p_k \quad (5)$$

3 Experiment Validation

3.1 Experiment Setup

To evaluate the performance of DLFi, we deploy a wireless sensor network to collect CSI measurements. We use a TL-WDR5300 wireless router with three antennas, operating in the 2.4 GHz band, as a transmitter. The receivers are Lenovo desktops running the Ubuntu 10.04LTS operating system, integrated with IWL5300 cards with three external antennas. Using the modified device driver enables the CSI measurements to be exported from the receivers. The placements of the AP and the MPs are fixed and known a priori, and both desktops can receive packets from AP concurrently.

Consider that environmental changes have impact on positioning performance, to verify the validity of DLFi method, we conducted our experiments in two different classical indoor environments: a laboratory, and a meeting room. As illustrated in

Fig. 5(a), the laboratory is a cluttered environment with many metal tables, chairs, and desktops blocking most of the LOS paths. Conversely, the meeting room is almost empty, so that most of the locations measured have LOS reception, which shown in Fig. 5(b).

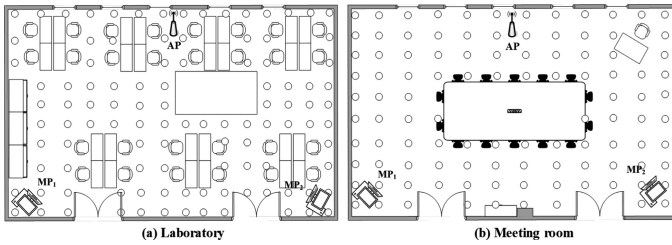


Fig. 5. Floor plans of different scenarios. (a) laboratory; (b) meeting room.

To reduce the effect of other factors on our technique, and to cover as large an area of interest as possible, the relative location of the devices is similar in both scenarios. The MPs are placed at the corners of one side of the laboratory and the meeting room, while the AP is placed at the center of the opposite side, as depicted in Fig. 5. This image also shows the reference points from which training data was collected as white circles in the free space of the two environments considered.

In the offline phase, CSI data are collected with the tester facing four different directions at each reference point. This reduces the magnitude of possible deviations between the similarity of test and training CSI datasets (caused by the differing orientations of a target in each respective phase), as in the test phase, there is no restriction on the orientation of the target during data collection. Similarly, location data for training are collected singly from each reference point, to ensure the presence of multiple targets in the area of interest does not affect the fingerprint database.

3.2 Performance of Intrusion Detection

We validate the performance of intrusion detection in the two different scenarios separately. To do this, we introduce two indicators, false positive (FP) and detection rate (DR), to characterize the performance of our technique. FP refers to the probability of erroneous detection, i.e., someone is detected when there is no person in the room, or no one is detected when there is someone in the room. In contrast, DR refers to the probability of accurate detection, i.e., detection is positive when someone is in the room and negative when the room is vacant.

Figure 6 shows the values of FP and DR in both scenarios considered. We observe that DLFi has the DR of over 90% and the FP of lower than 10% in both the laboratory and the meeting room, indicating that its use is feasible for intrusion detection. Hence, this detection result can be leveraged to improve the accuracy of localization.

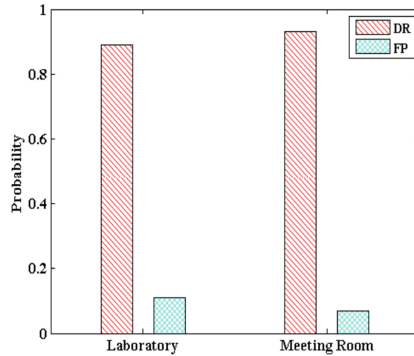


Fig. 6. False positive (FP) rate and detection rate (DR) in the two experimental scenarios considered.

3.3 Performance of Localization

To evaluate the performance of our technique in comparison to the Nuzzer (a device-free RSS-based indoor localization system) and Pilot (a two-stage device-free CSI-based indoor localization system) methods, we chose Nuzzer and Pilot for comparison because Nuzzer and Pilot are more classical methods in indoor positioning research based on RSS and CSI, respectively. Although these two methods were proposed in 2013 and do not represent the latest research level, they represent the most classic indoor positioning methods based on RSS and CSI, reflecting the basic level of research on indoor localization. Most research teams are improving on these basic researches to improve positioning performance, our team is no exception. Therefore, using these classic methods as references for performance analysis can more clearly explain what we have achieved. In addition, the experiment equipment and scenarios we used are very similar to those used in Nuzzer and Pilot, which reduce the interference of other factors, and improve the credibility of the experimental results. In summary, we use Nuzzer and Pilot for comparative analysis in the experimental comparison to illustrate the results achieved by DLFi.

We calculate the cumulative distribution function (CDF) of localization error from experiments conducted in the two representative indoor environments. The results of these calculations are illustrated below. Figure 7 depicts the CDF of localization error obtained in the laboratory and meeting room. In the Fig. 7(a), we note that DLFi is the most accurate of the three techniques considered, with localization error of 1 m for over 50% of the test points in the complex propagation environment (with tables obstructing most LOS paths and amplifying the multipath effect). And in the Fig. 7(b), with DLFi, over 70% of the test points produce errors of under 1 m, about 56% is similar to that obtained with Pilot. In contrast, with the Nuzzer method, only 33% of the test points produce errors lower than this minimum. We attribute this improved performance to the more detailed information implicit to the CSI measurements used in DLFi, compared to the RSS readings leveraged by Nuzzer. Furthermore, we also note improvements in comparison to Pilot, the other CSI-based localization system, demonstrating that the

use of only one correlation feature and a two-stage location classification method is less effective than the approach proposed in this paper.

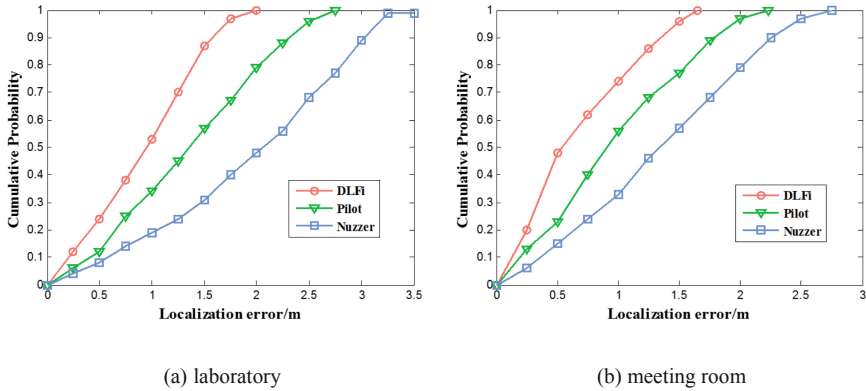


Fig. 7. Cumulative distribution function (CDF) of localization errors in two scenarios.

We conducted a further comparison between our technique and the Pilot and Nuzzer methods, based on mean error, standard error and localization accuracy. The results of this comparison are summarized in Table 1 below. From Table 1, we note that the mean error when DLFi is applied in a laboratory environment is 0.96 m, and the Pilot is 1.21 m, and Nuzzer is 1.86 m. In the meeting room, the mean error of DLFi is approximately 0.83 m, and the accuracy is 6.9% better than the Pilot method, and 13.9% better than the Nuzzer method.

Table 1. Statistical comparison of intrusion detection and localization performance in different experimental scenarios.

Method	Scenario	Mean error	Standard error	Accuracy
DLFi	Laboratory	0.96 m	0.87 m	89.4%
	Meeting Room	0.83 m	0.75 m	94.9%
Pilot	Laboratory	1.21 m	0.96 m	82.4%
	Meeting Room	0.98 m	1.08 m	88.8%
Nuzzer	Laboratory	1.86 m	1.25 m	79.1%
	Meeting Room	1.48 m	1.12 m	83.3%

Based on the comparison between the different localization techniques detailed above, it can be concluded that DLFi is capable of improving the accuracy of estimating a target's position. There are three reasons for these improvements. These are, the adoption of CSI as the data recorded in the fingerprint database, which contains more signal features, and has a finer granularity than RSS, pretreatment of the data

collected in the offline stage, which means that only the main characteristics of this data is retained; and the use of the results of intrusion detection to screen the fingerprint database during localization. This database screening is equivalent to two-phase positioning, without increasing the time cost. In addition, the kernel function used in our online positioning method reduces the computational complexity compared to the Pilot and Nuzzer, and processing is consequently quicker.

4 Parameter Optimization

4.1 Device Height

In validating the performance of our technique, we observed that the height of the devices and the tester affected the magnitude of the change in CSI measurements. Hence, we conducted a set of experiments to characterize this phenomenon, as the size of the variation in CSI measurements affects the ease of intrusion detection; large changes in CSI data make intrusion detection easier, while small changes make intrusion detection more difficult. In this set of experiments, the heights of the devices were set to 1 m, 1.5 m, and 1.9 m, while the heights of testers were 1.2 m and 1.8 m. Each tester stood in the same location while the heights of the devices were varied. As we observed varying changes to CSI, we utilize the probability of observing a change between the reference and measured CSI to illustrate the results of these experiments, as shown in Table 2.

Table 2. Comparison of the effect of heights on change in CSI measurements.

Height of devices	Height of testers	Probability of CSI change
1.0 m	1.2 m	90%
	1.8 m	85%
1.5 m	1.2 m	75%
	1.8 m	92%
1.9 m	1.2 m	70%
	1.8 m	78%

From Table 2 we note that when the devices are placed at a height of 1 m, the probability of observing a change in CSI measurements is large, regardless of the height of the testers. Conversely, when the devices are placed 1.9 m above ground, the probability of observing a change in CSI measurements is low. With the devices placed 1.5 m above the ground, the heights of the testers have a more significant effect on signal propagation; the probability of observing a change in CSI measurements was large with the 1.8-m tall tester, and small with the 1.2-m tall tester. From these results, we infer that the height of the subjects should be taller than that of the devices to ensure that the change in CSI measurements is noticeable. To validate this assumption, we repeated this experiment with a larger range of testers with varying heights, which confirmed our hypothesis. Hence, we kept all devices in the same plane, and set their

height to be shorter than that of the average tester, to prevent this from affecting the performance of our technique.

4.2 Number of Communication Links

As both the IWL 5300 card and the TL-WR740 N router have three antennas, there are multiple links between the AP and MPs. Although increasing the number of links means that more comprehensive signal characteristics can be obtained, and the size of the localization errors can be reduced, the increase in the amount of data increases the processing time. The effect of the number of links on localization error is shown in Fig. 8. Figure 8(a) illustrates the localization error in the laboratory and Fig. 8(b) shows the localization error in the meeting room. From this, we note that with the three techniques investigated, increasing the number of links decreases the magnitude of the localization error, with the worst errors being observed with a one transmitter-one receiver link (1TX-1RX) setup, and the best accuracy observed with a 2TX-3RX link setup, which is worse than others at the expense of processing time. In spite of this trade off, we conducted experiments using the 1TX-3RX link setup, to reduce the magnitude of localization errors.

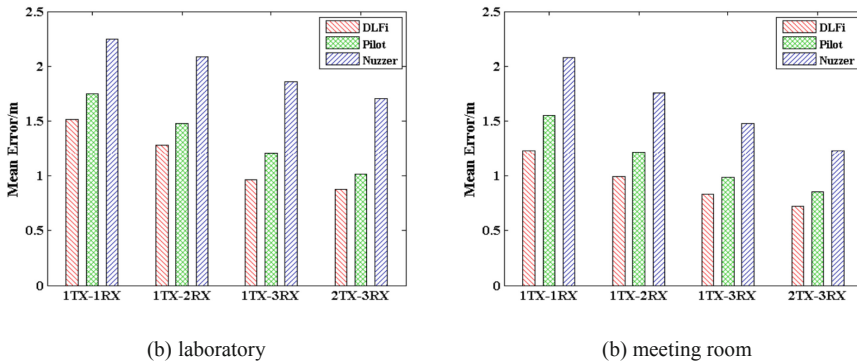


Fig. 8. Effect of the number of links on localization error.

4.3 Distance of Reference Points

To verify the effect of the distance between the reference points on localization error, we conducted further experiments where this parameter was varied. These experiments were conducted in the two scenarios defined previously, with the distances between the reference points set to 0.25 m, 0.5 m, 0.75 m, and 1 m. The results of these experiments are shown in Fig. 9. From Fig. 9, we note that localization error increases with the distance between the reference points, while a smaller distance leads to more accurate localization. However, reducing this distance also increases the number of reference points in the area of interest. Hence, more data is collected in building the fingerprint database, which makes processing more complex. In addition, reducing the

distance between these reference points does not have a constant effect on localization error; the improvements to positioning accuracy caused by reducing this distance from 0.75 m to 0.25 m, are minimal compared to those caused by reducing the distance from 1 m to 0.75 m. Therefore, an appropriate distance that minimizes localization error and ensures a manageable processing complexity should be chosen. For the scenarios discussed in this paper, this distance is set to 0.75 m.

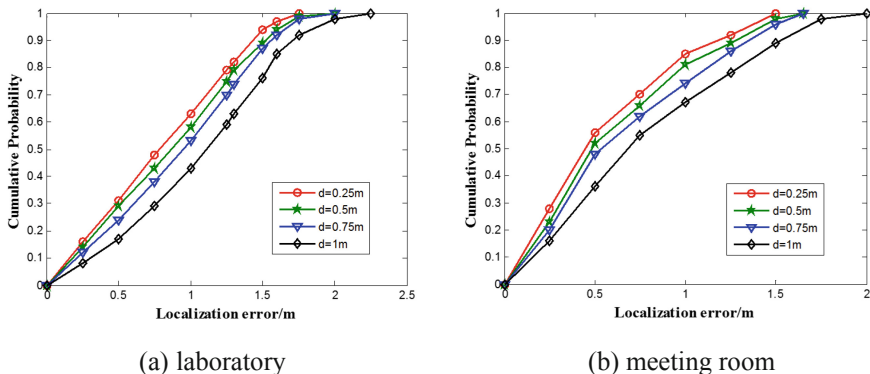


Fig. 9. CDF of localization error in different environments as a function of distance between reference points.

5 Conclusions

In this paper, we have presented an intrusion detection and localization method that makes use of CSI. In the offline phase of operation, CSI measurements are collected from two different receivers. As the redundancy of statistical variables in CSI is high, this raw data is subsequently processed using the PCA algorithm, which also reduces the magnitude of noise in these measurements. Following this, salient features of the CSI are extracted and stored in a fingerprint database. In the online phase of operation, intrusion detection is completed by comparing test data with the features stored in the fingerprint database, using the EMD algorithm. If a target is detected, the intruder’s probable location is determined according to the magnitude of the change in CSI data at different MPs. Based on these results, reference points are then selected, to build a sub-fingerprint database that effectively diminishes the area of interest. Finally, the precise position of the target is evaluated using the improved kNN algorithm, with weights calculated using the Gaussian kernel function. Our results demonstrate that employing CSI extracted from Wi-Fi improves the accuracy of intrusion detection and localization. A comparison of our technique with the Nuzzer and Pilot indoor localization methods illustrates its ability to reduce the magnitude of localization error, leading to improved positioning accuracy.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant 61762079, 61662070), Key Science and Technology Support Program of Gansu Province (Grant 1604FKCA097, 17YF1GA015), Science and Technology Innovation Project of Gansu Province (Grant CX2JA037, 17CX2JA039).

References

1. Shin, H., Chon, Y., Kim, Y., Cha, H.: MRI: Model-based radio interpolation for indoor war-walking. *IEEE Trans. Mob. Comput.* **14**(6), 1231–1244 (2015)
2. Lin, Y.-W., Lin, C.-Y.: An interactive real-time locating system based on bluetooth low-energy beacon network. *Sensors* **18**(5) (2018). Article ID 1637
3. Yang, B., Lei, Y., Yan, B.: Distributed multi-human location algorithm using naive Bayes classifier for a binary pyroelectric infrared sensor tracking system. *IEEE Sens. J.* **16**(1), 216–223 (2015)
4. Zhuo, R., Luo, L., Li, Z., Sang, N.: An indoor pedestrian position algorithm based on smartphone sensor. *Comput. Eng.* **42**(11), 22–26 (2016). (in Chinese)
5. Xiao, F., Wang, Z., Ye, N., Wang, R., Li, X.-Y.: One more tag enables fine-grained RFID localization and tracking. *IEEE/ACM Trans. Netw.* **26**(1), 161–174 (2018)
6. Husen, M.N., Lee, S.: Indoor location sensing with invariant Wi-Fi received signal strength fingerprinting. *Sensors* **16**(11) (2016). Article ID 1898
7. Pau, G., Collotta, M., Maniscalco, V., Choo, K.-K.R.: A fuzzy-PSO system for indoor localization based on visible light communications. *Soft. Comput.* **9**, 1–11 (2018)
8. Yasir, M., Ho, S.W., Vellambi, B.N.: Indoor positioning system using visible light and accelerometer. *J. Lightwave Technol.* **32**, 3306–3316 (2014)
9. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Predictable 802.11 packet delivery from wireless channel measurements. In: *SIGCOMM 2010: Proceedings of the ACM SIGCOMM 2010 Conference*, pp. 159–170 (2010)
10. Gao, Q., Wang, J., Ma, X., Feng, X., Wang, H.: CSI-based device-free wireless localization and activity recognition using radio image features. *IEEE Trans. Veh. Technol.* **66**(11), 10346–10356 (2017)
11. Bahl, P., Padmanabhan, V.N.: RADAR: an in-building RF-based user location and tracking system. In: *Proceedings. IEEE INFOCOM 2000 Conference on Computer Communications*, vol. 2, pp. 775–784 (2000)
12. Shi, S., Sigg, S., Chen, L., Ji, Y.: Accurate location tracking from CSI-based passive device-free probabilistic fingerprinting. *IEEE Trans. Veh. Technol.* **67**(6), 5217–5230 (2018)
13. Wang, X., Gao, L., Mao, S., Pandey, S.: CSI-based fingerprinting for indoor localization: a deep learning approach. *IEEE Trans. Veh. Technol.* **66**(1), 763–776 (2017)
14. Youssef, M., Agrawala, A.: The Horus location determination system. *Wirel. Netw.* **14**(3), 357–374 (2008)
15. Xiang, P., Ji, P., Zhang, D.: Enhance RSS-based indoor localization accuracy by leveraging environmental physical features. *Wirel. Commun. Mob. Comput.* **2018** (2018). Article ID 8956757
16. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Tool release: gathering 802.11n traces with channel state information. *ACM SIGCOMM Comput. Commun. Rev.* **41**(1), 53 (2011)
17. Zhou, R., Lu, X., Zhao, P., Chen, J.: Device-free presence detection and localization with SVM and CSI fingerprinting. *IEEE Sens. J.* **17**(23), 7990–7999 (2017)

18. Balog, M., Ilić, K., Mlinac-Jerkovic, K., et al.: Gender difference in glucocorticoid, insulin and estrogen receptors expression upon chronic stress and aging. In: RECOOP 13th Annual Scientific Conference, Bridges in Life Sciences (2018)
19. Wang, Y., Ma, X., Qian, P.: Wind turbine fault detection and identification through PCA-based optimal variable selection. *IEEE Trans. Sustain. Energy* **9**(4), 1627–1635 (2018)
20. Dhineshkumar, K., Subramani, C.: Kalman filter algorithm for mitigation of power system harmonics. *Int. J. Electr. Comput. Eng.* **8**(2), 771–779 (2018)
21. Zhou, R., Chen, J., Lu, X., Wu, J.: CSI fingerprinting with SVM regression to achieve device-free passive localization. In: 2017 IEEE 18th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–9 (2017)
22. Zhang, R., Zhang, J., Gao, Y., Zhao, H.: Block Bayesian matching pursuit based channel estimation for FDD massive MIMO system. *AEU-Int. J. Electron. Commun.* **93**, 296–304 (2018)
23. Wu, Z., Xu, Q., Li, J., et al.: Passive indoor localization based on CSI and Naive Bayes classification. *IEEE Trans. Syst. Man Cybern.: Syst.* **48**(9), 1566–1577 (2017)
24. Chen, H., Zhang, Y., Li, W., et al.: ConFi: convolutional neural networks based indoor wi-fi localization using channel state information. *IEEE Access* **PP**(99), 1 (2017)
25. Wang, X., Wang, X., Mao, S.: CiFi: deep convolutional neural networks for indoor localization with 5 GHz Wi-Fi. In: IEEE ICC 2017 - 2017 IEEE International Conference on Communications-Paris, France, 21 May 2017–25 May 2017, pp. 1–6. IEEE (2017)
26. Seifeldin, M., Saeed, A., Kosba, A.E., El-Keyi, A., Youssef, M.: Nuzzer: a large-scale device-free passive localization system for wireless environments. *IEEE Trans. Mob. Comput.* **12**(7), 1321–1334 (2013)
27. Xiao, J., Wu, K., Yi, Y., Wang, L., Ni, L.M.: Pilot: passive device-free indoor localization using channel state information. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems, pp. 236–245 (2013)



Wi-SD: A Human Motion Recognition Method Based on CSI Amplitude and Phase Information

Xiaochao Dang^{1,2}, Tong Zhang¹, Zhanjun Hao^{1,2(✉)}, and Yuexia Li¹

¹ College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, Gansu, China

zhanjunhao@126.com

² Gansu Province Internet of Things Engineering Research Center, Northwest Normal University, Lanzhou 730070, Gansu, China

Abstract. In the indoor environment, the monitoring of personnel activity behavior becomes more and more important. Although the traditional camera monitoring method has good performance, due to the limitation of deployment mode, there are monitoring blind spots and its deployment scope involves privacy issues. The non-device personnel acquisition and motion recognition through WiFi equipment, as a new type of highly promising technology, has received more attention and research. In this paper, we propose a human motion recognition method based on channel state information (CSI) amplitude phase mixing information, and classify the different activities of people. Different from the traditional single-person daily activity behavior recognition, this program focuses on the human exercise behavior of different people with different intensity, and promotes to the related sports behavior recognition of two people. Compared with the single person situation, the strength, amplitude and regularity of the two people exercising at the same time are very different. We experimentally tested the effects of different activities of single and double on CSI in two real environments, extracted relevant amplitude and phase information, and used machine learning to summarize the change patterns classification. At the same time, consideration of the line-of-sight factor has improved the overall flexibility of the system and improved the condition of motion recognition.

Keywords: Human motion · Channel state information (CSI) · Support vector machine (SVM) · Dynamic time warping (DTW)

1 Introduction

With the development of communication technology, indoor human behavior recognition technology has become a new and promising research direction and has achieved fruitful research results. Human behavior recognition has mature research including intrusion detection, fall detection, gesture recognition, gait recognition, motion analysis, etc., and is applied to the analysis of the health status of the elderly, the detection of accidental falls, and the safety monitoring of important places.

The traditional human activity recognition technology often needs to be worn by the detected object [1]. For such problems, the device-free passive (Dfp) behavior detection technology has higher flexibility and universality, can be continuously obtained indoors without being affected by the presence of personnel. Information greatly improves the convenience of human detection [2].

Currently, device-free passive human detection includes video based [3], infrared based signal [4], radio based signal [5], the human behavior activity sensing technology based on WiFi signal reflects its own unique advantages [6]. The two most commonly used feature signals in the field of wireless sensing are the Received Signal Strength Indicator (RSSI) [7] and the Channel State Information (CSI) [8]. The early WiFi-aware system uses the RSSI signal from the MAC layer to implement the CSI, and the proposed CSI as the physical layer information can better reflect the fine-grained features in the signal transmission process. CSI signals are affected by multipath effects during indoor propagation, scattering and reflection occur, and object motion, especially human activities, has a more significant effect. In the research of this paper, we not only consider the distinction and identification of single daily behaviors and severe abnormal activities, but also promote the analysis of activity behavior in the case of two people. This paper proposes a Wi-SD detection method for indoor human activity based on CSI amplitude and phase mixed signals, which is based on the differentiation of behaviors of different motion scales, and further realizes the specific activity behavior of root fine-grained. Judgment and identification. The Wi-SD method extracts the relevant feature information by acquiring the CSI signal under different behaviors of the personnel, and based on the energy change generated by the CSI phase information, uses the support vector machine (SVM) [9] to perform the initial classification process according to the severity of the human motion. And then using the CSI amplitude information according to the discriminating result, using the dynamic time warping (DTW) method [10] for further processing, by comparing the time domain feature information in the specific action, matching according to different subcarrier fingerprint differences, obtaining the current personnel Specific event information.

In summary, the contributions of our work are listed as follows:

We constructed a relationship model between human motion behavior and CSI signal, and based on the extracted data features, using machine learning methods to deal with different motion behaviors of indoor people.

We have studied the effects of different intense human activities on wireless signals, and considered the different signal characteristics generated by the interaction between two people in the case of two people. A behavior discrimination method based on CSI phase difference and amplitude is proposed. Creatively use the different characteristic information contained in the two, and carry out rough classification from the intensity of the sports behavior in stages, and then match the fine-grained behavior characteristics to achieve accurate identification of specific human motion behaviors.

The scheme was deployed on a TP-link Wi-Fi router equipped with OpenWrt system, and the control experiment was carried out under different environments. The influence of different environmental factors and obstacle interference on the experimental results was tested, and the system under LOS and NLOS conditions was evaluated stability and reliability.

The rest of the paper is organized as follows: In the second part, we will summarize the relevant work and technical principles. In the third part, we introduce the Wi-SD system architecture. In the fourth section, we introduce the effects of different degrees of motion and the interaction of two people on wireless signals. The fine structure and design method are described in five sections. In the sixth section, we conducted a test evaluation of the Wi-SD system through experimental verification. In the last section, we summarize the work of this article.

2 Related Work

In this section, we will summarize the existing methods of human behavioral activities. The existing device-free passive human behavior detection is roughly divided into two types: systems based on visual image recognition and systems based on wireless signals. The human body signal can be directly captured by the visual sensor for graphic image processing. At the same time, the interference of human activity on signal propagation can extract corresponding feature information and analyze the specific behavior.

In recent years, research on human motion detection systems based on WiFi has become more and more mature. This technology has been widely used in practice, including gait detection [11], gesture recognition [12], sleep monitoring [13], trajectory tracking [14] and so on. It is mainly divided into two sub-categories based on perceptual signals, RSSI-based sensing systems and CSI-based sensing systems.

RSSI-based: RSS is a WiFi-based signal strength indicator. It was used in the early construction of indoor positioning system [15, 16]. By analyzing the influence mode of human movement on signal generation, the relationship model between location and RSS is constructed to realize indoor positioning. Further research found that using the variance information of RSS can perform simple motion detection, and the proposed WiSee system can perform gesture recognition monitoring without equipment. However, as the information extracted from the MAC layer, RSS is used to measure the strength information of the received data frames. It has good performance when detecting large-scale coarse-grained operations, and it is difficult to perform more fine-grained human motion recognition.

CSI-based: The channel state information CSI describes the phase and amplitude information carried by each subcarrier as information extracted from the physical layer, which better reflects the true trend of the signal. At the same time, each subcarrier, especially its independence and difference, can reflect more fine-grained physical motion information. In [17], the E-eyes system proposed by Wang et al. In 2014, by analyzing the amplitude distribution of CSI signals, identified 11 kinds of fixed space movements including washing dishes and cooking, and walking from bedroom to kitchen. The spatial action focuses on the analysts who have different behaviors in a single person situation. In the literature [18], the author proposes a detection scheme CareFi for sedentary daily office behavior, which has a good performance in judging small-scale daily fine-grained behavior.

Inspired by the above research, we propose a detection system for identifying the specific movement state of indoor personnel. This paper not only discusses the

identification and monitoring of sudden behavioral actions in the case of single and double, but also considers the violent activities that distinguish abnormalities. In the case of further testing of the behavior of the personnel.

3 System Overview

Based on the correlation between different human motion and CSI feature information, we propose a quadratic classification method for complex human motion behavior identification Wi-SD. The whole framework is shown in Fig. 1, which includes data preprocessing, feature library establishment, rough classification of motion behavior, and specific action recognition. First, our system extracts CSI data and preprocesses the data using outlier removal and smoothing noise reduction techniques to reduce interference. Monitor the CSI stream and use the coarse identification method to classify the current state as a dynamic or static activity. Then, after the noise is removed, different categories are processed by different recognition methods. According to the intensity of exercise, firstly, the human body movements are roughly classified. In the system, we choose to use SVC to classify the extracted CSI phase difference information for the first time to determine the number of people currently exercising and the intensity of exercise; then use the dynamic time. The regularization technology DTW then classifies the CSI amplitude time domain information twice, so as to accurately determine the current specific motion behavior content. This system is used as an experimental basis for subsequent research.

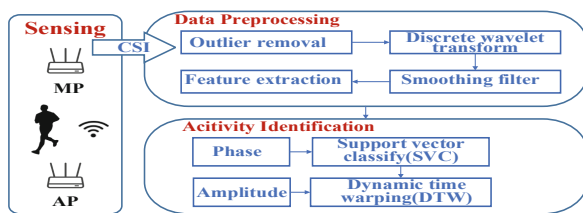


Fig. 1. Wi-SD system framework

4 CSI-Based Human Motion Detection

4.1 Channel State Information

CSI describes the reflection, diffraction and scattering that a signal undergoes during propagation. Current commercial wireless devices employ Orthogonal Frequency Division Multiplexing (OFDM) at the physical layer and comply with the IEEE 802.11n/ac standard, allowing multiple transmit and receive antennas for multiple input, multiple output (MIMO) communications. CSI combines the time delay of multiple paths on each subcarrier, the effect of amplitude attenuation and phase shift. CSI is a description of the attenuation factor experienced in signal transmission and is

an estimate of the gain matrix. Then the subcarrier formula of a single CSI signal is as follows:

$$H_i = |H_i|e^{j\sin\theta} \tag{1}$$

The θ is the subcarrier phase, and H_i is the subcarrier amplitude. Under the IEEE 802.11n protocol, the bandwidth affects the number of subcarriers. When the bandwidth is 20 MHz, the number of subcarriers in a single group is 56. In this paper, CSI information is obtained through a commercial network card. Each CSI signal represents a matrix information of $3 \times 2 \times 56$, where 3 is the number of receiving antennas, 2 is the number of transmitting antennas, and CSI data contains time delays of multiple paths. Human activity is a continuous action, which is reflected in continuous continuous changes in the signal time domain, and can extract physical features of corresponding features and specific actions. WiFi wireless signals can be modeled as channel impulse response (CIR) in the time domain, and the expression of $h(t)$ is

$$h(t) = \sum_{l=1}^L \alpha_l e^{j\phi l} \delta(t - t_l) \tag{2}$$

α and ϕ correspond to amplitude and phase under different multipath components, respectively, t_l is time delay, L is total number of multipaths, $\delta(t)$ is dicla function. The CSI time domain information contains the propagation delay and Doppler shift information generated by continuous environmental changes. The following Fig. 2 reflects the changes in signal transmission caused by human actions and movements.

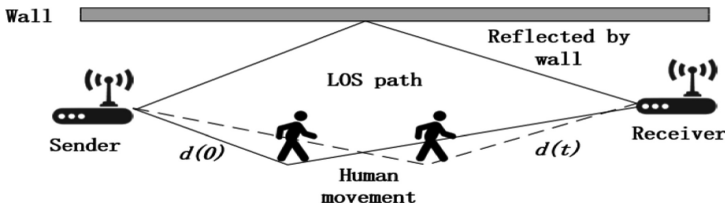


Fig. 2. Propagation model of human motion CSI

It can be seen from the above model that the energy attenuation and propagation delay generated by multi-channel characteristics can reflect more complex human behavior change rules in a finer granularity. By extracting CSI data under different motion states, selecting amplitude and phase difference to construct timing. Signal model, based on this analysis of human motion characteristics.

4.2 The Effect of Human Motion on CSI Amplitude

Different human activities are affected by conditions such as the range of motion, the frequency of movement, and the number of people exercising, which will produce

different signal characteristics. In order to analyze the more complex human motion states, we choose to extract the CSI data features generated by various motion behaviors through experiments. Find the mapping between human motion and signals.

We first observe the effects of different behaviors on CSI signals in a single person situation. Figure 3 shows the change of CSI signal amplitude when two different motion behaviors are carried out and running. In order to facilitate the data preprocessing process, the single group image is the original amplitude information from top to bottom, after wavelet transform. Amplitude information and smoothed amplitude information.

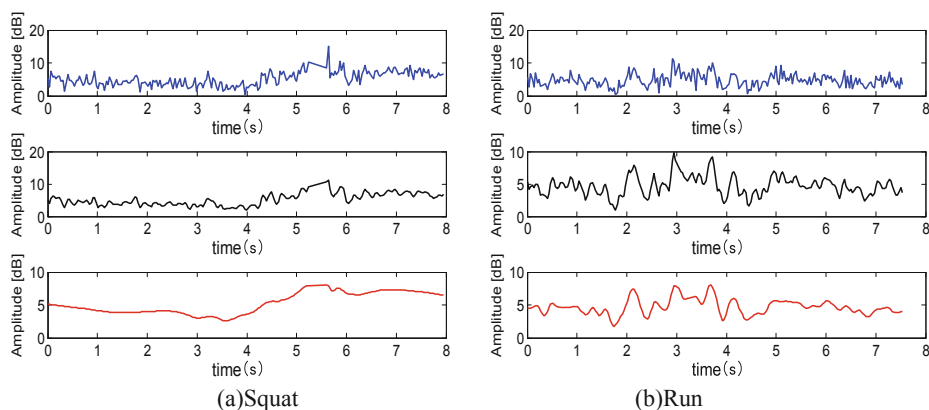


Fig. 3. Time domain map of human motion CSI amplitude

It can be seen from Fig. 3 that in the case of single person, the amplitude information reflected by different motion changes has obvious differences, which can be used as the discriminating basis for different actions. On this basis, we conducted an experiment of two-synchronous motion for comparison. The difference caused by the human situation. Figure 4 is the amplitude information of single running and double running at the same time.

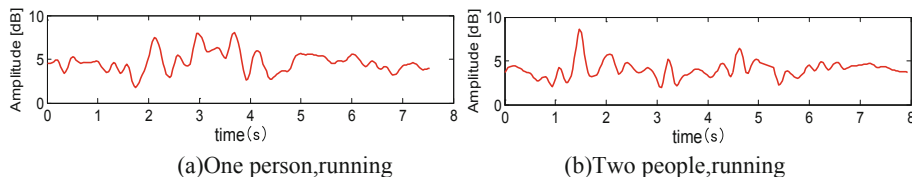


Fig. 4. CSI amplitude map for different numbers of people

When the two people perform the synchronous action, although the peak change tendency similar to the single person situation is maintained, the overall amplitude fluctuation tendency increases. Through experimental comparison, it is found that when the number of people is consistent with the intensity of exercise, the amplitude characteristics can map different behaviors well, which is used as the criterion for human motion judgment. However, in the case where the two conditions are different and the motion state is relatively strong, the corresponding amplitude information generated is not sufficiently high, and the feature information extracted on the existing time domain map is difficult to generate. Certainly confused. It is difficult to accurately identify complex situations by simply using CSI amplitude information. Therefore, it is necessary to first perform rough classification before determining the specific motion, and judge the severity of the current motion. Based on the rough classification, further realize the specific motion determination.

4.3 Complex Motion and CSI Changes

In order to study the energy attenuation of signals generated by different severe motion behaviors, we introduce phase difference information to analyze the influence of the intensity of motion states on CSI signals. The phase difference expression is as follows:

$$\Delta\hat{\phi}_i = \Delta\phi_i + 2\pi f_i e + \Delta\beta + \Delta Z \quad (3)$$

Where $\Delta\hat{\phi}_i$ is the true phase difference, e is the time lag difference between the antennas, $\Delta\beta$ is the unknown phase offset, and ΔZ is the noise. The randomly distributed original phase can be calibrated by phase difference, which is affected by environmental and human motion. Figure 5 shows the phase difference image for single walking, single running and double running.

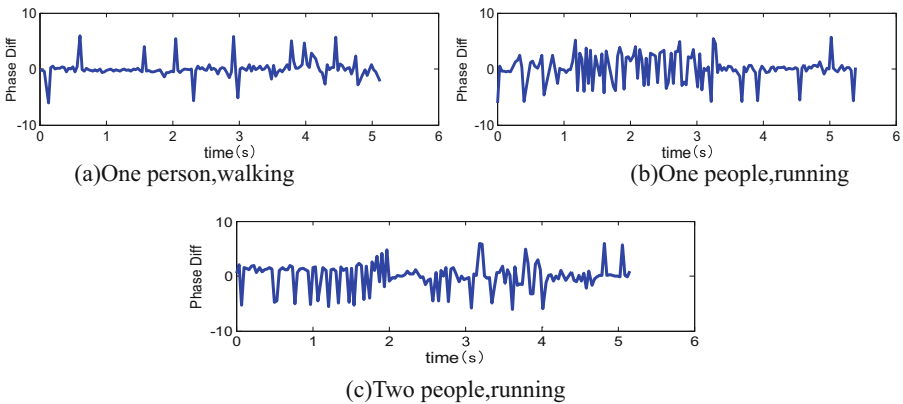


Fig. 5. CSI phase difference diagram of human motion

It can be seen from Fig. 5 that the change of signal energy caused by the increase of the target number will have a certain impact, but when the amplitude of the human body increases, the phase difference will change significantly. Compared with the amplitude information, the phase difference has a greater degree of reflection. Obvious features, which can be used to distinguish between different levels of exercise.

4.4 Summary

It is found in the study that the energy attenuation of the signal produced by different severe levels of motion is significantly better than the amplitude information in phase. However, since the phase difference itself is expressed as continuous peak fluctuation, it is impossible to extract sufficiently accurate features to perform specific motion discrimination. Therefore, we combine the phase difference and the amplitude to make a second discrimination on the human motion, and pass the phase. The difference is based on the initial classification of the intensity of exercise and the number of athletes. On this basis, the amplitude information is used for specific motion recognition.

5 CSI Feature Processing Method

5.1 Pretreatment

Before the data is officially used, due to environmental noise and equipment factors, abnormal measurements other than the effects of human motion will occur. We need to eliminate the obvious abnormal values and minimize the noise impact of the data itself. Studies have shown that the signal range caused by normal human motion is in the range of 0–5 Hz. Therefore, the threshold is selected to filter the uncorrelated signal components. In addition, in the system, smoothing filter is also used to take the tie value of the adjacent continuous data points, so that the overall data trend can obtain more obvious image features without affecting the feature information carried by the whole device, for subsequent extraction. The characteristics of motor behavior are facilitated.

5.2 SVM Classifier

For the rough classification problem of simple type samples, it is more efficient to select support vector machines for processing. According to the signal phase difference characteristics generated by different degrees of motion in different situations, establish a relevant rough mapping model.

Let q be the number of training samples, and construct training sample (k_i, g_i) , where k_i is the pre-processed each action feature sample data set, and g_i is the sample classification label. The SVC process is known as the sample set (k_i, g_i) . To find the most classified hyperplane, the SVC classification constructor is established as follows:

$$\begin{cases} \min(\frac{1}{2}\|w\|^2 + C \sum_{i=1}^q \eta_i) \\ s.t. \eta_i \geq 0 \\ g_i(w^T k_i + b) \geq 1 - \eta_i \\ C > 0 \end{cases} \quad (4)$$

Where w is the direction vector separating the hyperplane, b is the hyperplane position constant, C is the penalty parameter, η_i is the error. Solving the equations according to the constraints produces a classification function as follows:

$$f(k) = \text{sign}\left(\sum_{i=1}^l \alpha_i K(k_i, k) + b\right) \quad (5)$$

Where $K(k_i, k)$ is the kernel function that maps the CSI fingerprint to a higher dimension, and the radial basis function is selected as the kernel function, then $K(k_i, k_j) = \exp(-\|k_i - k_j\|^2)$. The position data in the fingerprint database is input as a training sample, and the Eq. (4) is linearly solved to obtain $\alpha_i = (\alpha_1, \alpha_2, \dots, \alpha_q)^T$ and b , which provides parameter support for the decision function used in online matching. We mark the normal daily exercise as a negative sample, mark the severe abnormal motion as a positive sample, determine the type of action currently being performed according to the classification function, and realize the initial detection of the human action type.

5.3 Dynamic Time Warping

On the basis of the type of exercise that has been obtained, it is necessary to further classify the sample features to determine the specific action behavior. In this system, we chose to use Dynamic Time Warping (DTW) to align the measured CSI amplitude time domain information with known sample profile data. We chose to use the amplitude similarity information of the DTW to compare the appropriate values between the two sequences to match the sample configuration corresponding to the peak information generated by different actions. We build multidimensional DTW based on multiple subcarrier information of CSI data. The formula is as follows:

$$d(c_i, c'_h) = \sum_{n=1}^N (c_i(n) - c'_h(n))^2 \quad (6)$$

c and c' are the sample sequence and the test sequence, respectively, and p is the matrix dimension. Through the above formula, the focus is on finding the lowest cost path, and determining the sum of the path of each element. According to the principle of minimization, the amplitude information is matched and Minimum measurement path results to determine the sample to which the action belongs, thereby enabling identification of specific actions.

6 Evaluation

6.1 Experimental Setup

In the experiment of this system design, two sets of built-in Atheros AR9580 NICs and TPLink WDR4310 routers equipped with Openwrt system were selected, and the above two machines were set as transmitter MP and receiver AP respectively. Our platform is capable of recording complete CSI data for 114 subcarriers using 5 GHz communication with 40 MHz bandwidth. Fine-grained CSI reflects more precise movements and environmental changes in humans.

The experimental scene selects the laboratory and the conference room respectively. In two different scenarios, the individual daily actions such as walking, bending, standing up, picking up, waving, and single-person strenuous actions such as running, falling, etc. are tested experimentally. At the same time, on the basis of this, choose double to repeat the above-mentioned actions, and compare the stability of the system under different numbers of people. The actual scene of the experimental site is shown in Fig. 6.

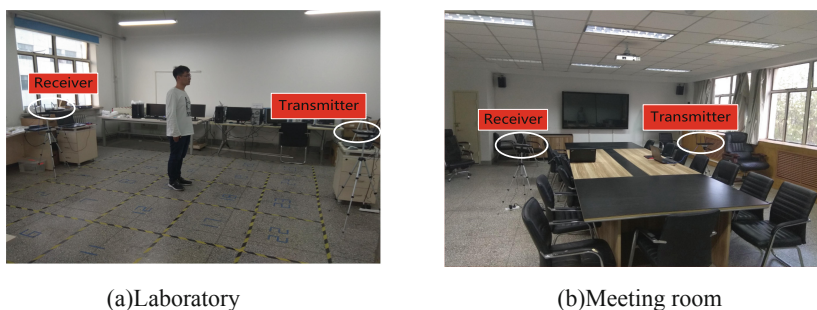


Fig. 6. Experimental scene

It can be seen that the laboratory layout is relatively compact, there are more devices, and multipath interference is stronger; the layout of the conference room is simple and the whole is relatively empty; by adjusting the relative positions of the transmitter and the receiver, it can be tested under LOS and NLOS conditions system performance difference.

6.2 Performance Analysis

Specific Action Recognition Rate in Different Environments. The focus of this system is to detect different human body movements. Therefore, we choose to detect different actions in stages. The test models are mainly divided into two categories, one is daily behavioral actions, such as standing, walking, etc., which are less harmful to signal infection; the other is abnormally vigorous exercise, such as fighting, falling, running, etc. Interference behavior. The above experiments were first designed to be

tested in two different experimental scenarios in the laboratory and conference room under single-person conditions. A total of seven groups of actions were performed. To ensure the stability of the test results, the testers and experimental equipment were selected to be unified. Experimental scenario is a variable condition. The experimental results are shown in Fig. 7.

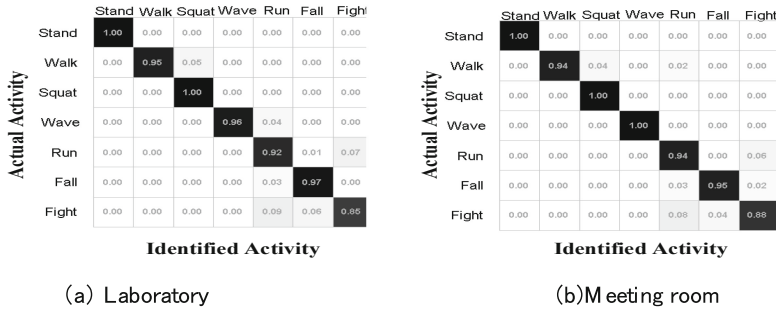


Fig. 7. Motion recognition

It can be seen from Fig. 7 that in the two cases of the laboratory and the conference room, the overall recognition effect is similar, and the multi-path effect is less affected in the conference room environment, and the motion recognition accuracy is slightly higher. According to the results analysis, compared with the strenuous exercise situation, the system maintains higher recognition accuracy in daily motion recognition.

Test Population Impact on the System. We have tested the identification of different human movements in a single-person situation. In this section, we further consider the influence of the number of factors on the overall discriminating power of the system, and choose to test the seven movements in a single experimental environment. The results of the experiment are shown in Fig. 8.

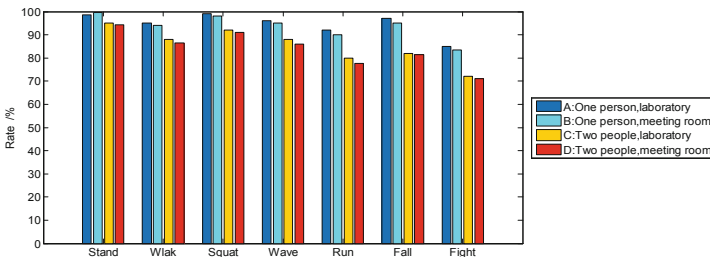


Fig. 8. Double situation test

It can be seen from Fig. 8 that in the case of controlling environmental conditions and action constraints, the recognition effect of single action is significantly higher than that of the same action performed by two people, and the recognition rate of daily actions such as standing, walking, bending, waving, etc. is significantly higher than that. Striking and complex movements of the human body such as running, falling, and fighting. After comparing the difference between the same action and the single person in the same situation, it can be found that the accuracy of the double recognition of strenuous exercise behavior is significantly lower than the single-person condition by more than 10%. The daily behavior is under the condition of two people, and the contrast difference is within 3% to 7%. From the above conclusions, compared with daily behavior, the increase in the number of people will make the recognition of complex movements significantly increase, reducing the accuracy of motion recognition.

LOS and NLOS Identification. In order to verify the validity of motion recognition in the case of LOS/NLOS, we set the corresponding conditions in the above two scenarios for experimental verification. In the experiment, we chose to place a metal plate of size $2\text{ m} \times 1.5\text{ m}$ as an obstruction between the transmitter and the receiver to create NLOS conditions. The whole experiment was carried out in different scenarios with LOS conditions and NLOS conditions for human behavior detection. According to the intensity of exercise, it is divided into two sets of reference data sets, and the following two metrics are introduced: (1) true positive rate (TPR) is defined as the percentage of correct detection of daily human behavior; (2) true negative rate (TNR) is defined as the percentage of correct abnormal motion behavior detected. For the convenience of evaluation, when inputting the sample set, it is divided into three categories: unmarked input and random selection of positive and negative samples as test set; labeled as LOS and random sample selected as LOS test set; labeled as NLOS and selected random sample as NLOS test set. The experimental results are shown in Table 1 and Fig. 9.

Table 1. Recognition rate of each action under LOS/NLOS conditions in different environments

Environment	Condition	Stand	Walk	Wave	Run	Fall
Laboratory	LOS	0.99	0.94	0.96	0.92	0.96
	NLOS	0.96	0.91	0.93	0.85	0.91
Meeting room	LOS	0.99	0.95	0.95	0.91	0.95
	NLOS	0.97	0.93	0.91	0.83	0.89

It can be seen from Fig. 9 that the deviation distribution of the LOS condition is more negative than the deviation distribution of the NLOS condition, and the LOS overall exhibits a higher detection efficiency than the NLOS, and the experimental scene influence is within a reasonable range. The effect of multipath effects caused by environmental congestion on NLOS is not sufficiently significant compared to LOS.

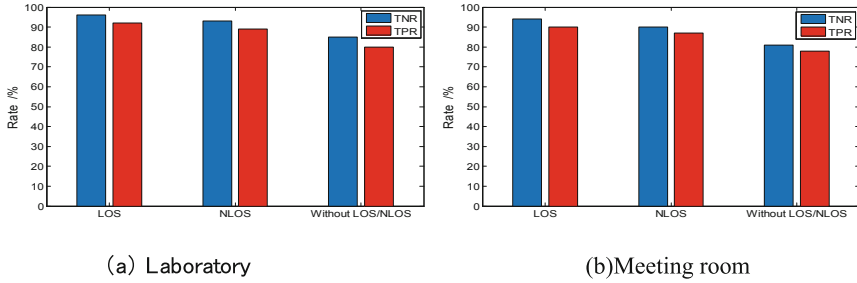


Fig. 9. LOS/NLOS evaluation

Algorithm Robustness Test. In practical applications, we need to consider the impact of the device’s own parameter adjustment on the system. The packet delivery rate of the wireless device directly affects the construction quality of the feature database and the efficiency of the algorithm. Considering the discrimination of the calculation stability, we choose to adjust the built-in parameters of the device in the two experimental environments for 10, 30, 50, 70, 100, 120 packet rate experiment, statistical analysis of algorithm execution time and human motion recognition accuracy rate in the system, sample number experiment and link number experiment result shown in Fig. 10.

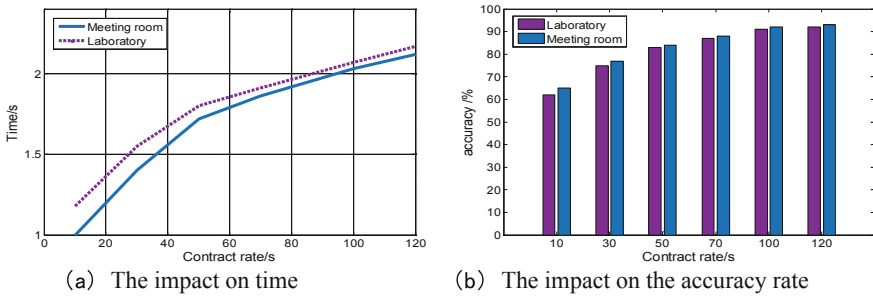


Fig. 10. The impact of the rate of delivery on the algorithm

Analysis of Fig. 10 shows that increasing the number of samples in both environments will improve the positioning accuracy and increase the execution time of the algorithm, which will affect the algorithm execution efficiency. In the conference room environment, the algorithm execution speed and recognition rate are slightly higher than the laboratory environment, and as the number of samples increases, the numerical difference between the two environments decreases. When the number of samples is between 10 and 50 packets per second, the recognition rate of the human action and the algorithm time continue to rise; when the number of samples is between 50 and 120 packets, the accuracy of the motion recognition is basically kept at the same level and tends to be stable. However, execution time is still rising. Based on the above results, it can be inferred that the selected packet rate is maintained at 50 to 70 packets per second to maintain algorithm efficiency while maintaining high motion recognition accuracy.

Comparison of Performance of Different Algorithms. The above experiments evaluated the effectiveness of the Wi-SD algorithm itself. In this section, we evaluate the accuracy of behavior recognition under different sampling numbers by comparing the other three algorithms. At the same time, we consider that when the test conditions are fixed, multiple sets of actions are performed continuously, and the number of action groups in the test set also affects the stability of the system. Here, the E-eyes and Care-Fi methods are respectively compared with the Wi-SD method, and the performance of the three methods is tested by setting different size training sets and setting the number of consecutive action groups. The result is shown in Fig. 11.

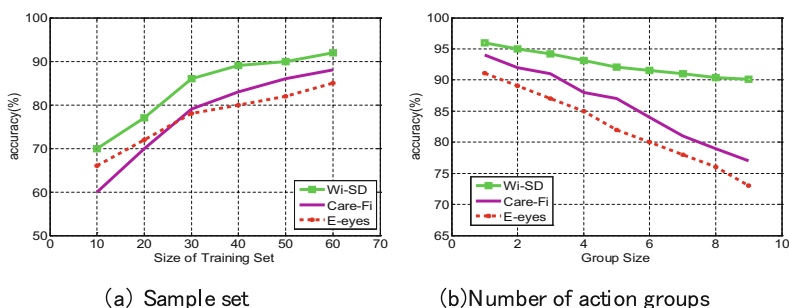


Fig. 11. Performance comparison of different methods

It can be seen from Fig. 11(a) that all three methods are positively correlated with the test integration, and in general, the Wi-SD algorithm is superior to the Care-Fi method and the E-eyes method when the number of test sets is greater than 50. The recognition accuracy is above 90%. As shown in Fig. 11(b), as the group size increases, the accuracy of Wi-SD decreases from 96% to 90%, and then tends to be stable. The accuracy of the other two methods is more than 4 when the number of continuous action groups exceeds 4. The decline is obvious. Therefore, the comparison results show that the method will run more stably when the group size becomes larger.

7 Conclusions

This paper proposes a two-stage complex human motion detection method based on CSI. Through the secondary classification method, it can accurately identify multiple actions. By analyzing the characteristics of different types of human movements, we can find out the differences and unique patterns of change between daily behaviors and strenuous behaviors. Through a large number of experiments, we tested the classification of motions with different severity and the recognition performance of the two people, and considered the difference between LOS and NLOS conditions. The experimental results show that the method has high stability under different environments and accurately identifies different human motion behaviors. In the future

research, we will further consider the impact of multi-person interaction on human behavior perception, and shift the focus from traditional motion recognition to complex multi-person concurrent perception, further deepening research.

References

1. Youssef, M., Mah, M., Agrawala, A.: Challenges: device-free passive localization for wireless environments. In: International Conference on Mobile Computing and Networking, pp. 222–229. ACM (2007)
2. Chetty, K., Smith, G.E., Woodbridge, K.: through-the-wall sensing of personnel using passive bistatic WiFi radar at standoff distances. *IEEE Trans. Geosci. Rem. Sens.* **50**(4), 1218–1226 (2012)
3. Aggarwal, J.K., Ryoo, M.S.: Human activity analysis: a review. *ACM Comput. Surv.* **43**(3), 1–43 (2011)
4. Rokade-Shinde, R., Sonawane, J.: Dynamic hand gesture recognition. In: International Conference on Signal and Information Processing (IConSIP), pp. 1–4. IEEE (2016)
5. Huang, X., Dai, M.: Indoor device-free activity recognition based on radio signal. *IEEE Trans. Veh. Technol.* **PP**(99), 1 (2017)
6. Pu, Q., Gupta, S., Gollakota, S., et al.: Whole-home gesture recognition using wireless signals. In: ACM SIGCOMM Conference on SIGCOMM, pp. 27–38 (2013)
7. Seifeldin, M., Saeed, A., Kosba, A.E., et al.: Nuzzer: a large-scale device-free passive localization system for wireless environments. *IEEE Trans. Mob. Comput.* **12**(7), 1321–1334 (2013)
8. Wu, C., Yang, Z., Zhou, Z., et al.: Non-invasive detection of moving and stationary human with WiFi. *IEEE J. Sel. Areas Commun.* **33**(11), 2329–2342 (2015)
9. Zhang, D., Wang, H., Wang, Y., Ma, J.: Anti-fall: a non-intrusive and real-time fall detector leveraging CSI from commodity WiFi devices. In: Geissbühler, A., Demongeot, J., Mokhtari, M., Abdulrazak, B., Aloulou, H. (eds.) Inclusive Smart Cities and e-Health. ICOST 2015. Lecture Notes in Computer Science, vol. 9102, pp. 181–193. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-19312-0_15
10. Zhang, O., Srinivasan, K.: Mudra: user-friendly fine-grained gesture recognition using WiFi signals. In: International on Conference on emerging Networking Experiments and Technologies, pp. 83–96(2016)
11. Yan, L., Zhu, T.: Using Wi-Fi signals to characterize human gait for identification and activity monitoring. In: IEEE First International Conference on Connected Health: Applications, Washington, USA, pp. 238–247 (2016)
12. Hong, L., Wei, Y., Wang, J., et al.: WiFinger: talk to your smart devices with finger-grained gesture. In: ACM International Joint Conference on Pervasive & Ubiquitous Computing, New York, USA, pp. 250–261 (2016)
13. Liu, X., Cao, J., Tang, S., et al.: Wi-sleep: contactless sleep monitoring via WiFi signals. In: Real-time Systems Symposium, Rome, Italy, pp. 346–355 (2014)
14. Joshi, K., Bharadia, D., Kotaru, M., et al.: WiDeo: fine-grained device-free motion tracing using RF backscatter. In: USENIX Conference on Networked Systems Design & Implementation, Oakland, USA, pp. 317–329 (2015)
15. Wang, X., Gao, L., Mao, S., et al.: CSI-based fingerprinting for indoor localization: a deep learning approach. *IEEE Trans. Veh. Technol.* **66**(1), 763–776 (2016)

16. Zhen, H., Luo, Z., Chen, Z., et al.: Indoor location algorithm based on optimizing least square support vector machine. *Acta Scientiarum Naturalium Universitatis Sunyatseni* **55** (02), 48–51 (2016)
17. Yan, W., Jian, L., Chen, Y., et al.: E-eyes: device-free location-oriented activity identification using fine-grained WiFi signatures. In: *International Conference on Mobile Computing & Networking*, New York, USA, pp. 617–628 (2014)
18. Yang, J., Han, Z., Hao, J., et al.: CareFi: sedentary behavior monitoring system via commodity WiFi infrastructures. *IEEE Trans. Veh. Technol.* **67**(8), 7620–7629 (2018)



Data-Quality-Aware Participant Selection Mechanism for Mobile Crowdsensing

Hongbin Sun and Dan Tao^(✉)

School of Electronic and Information Engineering, Beijing Jiaotong University,
Beijing 100044, China
dtao@bjtu.edu.cn

Abstract. Data quality assurance is one of the most critical challenges in the context of Mobile CrowdSensing (MCS). How to effectively select appropriate participants from large-scale candidates to perform sensing tasks while satisfying certain constraint is a problem to be solved. Motivated by this, this paper studies the problem of data-quality-aware participant selection for MCS. Firstly, we propose a quality-aware participant reputation model by introducing active factor to lay a theoretical foundation. Secondly, we present a Multi-Stage Decision solution based on Greedy strategy (MSD-G) to optimize the pending problem while satisfying certain data quality constraint. Extensive simulations over a real dataset verify that our proposed MSD-G can effectively realize participant selection with ideal recruitment cost and sensing data quality.

Keywords: Mobile CrowdSensing · Data quality · Reputation model · Participant selection

1 Introduction

Mobile CrowdSensing (MCS) is a new paradigm of applications that utilizes ubiquitous mobile devices to collect and share sensing data from surrounding environment over a large geographical region [1]. Compared to traditional static sensor networks, MCS has distinct advantages, such as low-cost deployment & maintenance, flexible mobility. However, some subjective factors (e.g., willingness, malicious behavior) in data collection process and objective factors (e.g., professional skills, device performance, environment) may greatly affect sensing data quality, user reputation and participant selection [2, 3]. It is obvious that continuous low-quality data will do harm to the service credibility of a sensing platform.

The assurance of sensing data quality brings new challenges. One of the key challenge is to motivate participants to collect high quality data while constraining participants' malicious sensing behavior. The existing research findings focused on incentive mechanisms [4, 5], which were crucial for the recruitment of mobile users to participate in a sensing task and to ensure that participants provide high-quality sensing data. Also, recent research has shown that reputation based schema can be useful for

accomplishing sensing tasks with high quality and low cost (e.g., reward) [6]. Another key challenge is participant selection problem, that is, how to effectively select appropriate participants from large-scale candidates to perform sensing task while satisfying certain constraints. For a sensing task in MCS, the more reliable participants are involved, the better spatio-temporal coverage achieves, and the less sensing cost paid by data requester will become. In fact, the sensing cost is proportional to the scale of participant scale once the basic reward per participant is fixed. Recently, some studies [7–9] have addressed this problem. *Pournajaf et al.* [7] examined the problem of spatial task assignment in crowd sensing when participants utilized spatial cloaking to obfuscate their locations. However, they merely considered the spatial tasks while ignoring temporal requirements. *Liu et al.* [8] presented a QoI-aware energy-efficient participant selection approach to provide a suboptimal solution to the defined optimization problem. *Zhang et al.* [9] proposed a participant selection framework, named CrowdRecruiter, which minimized incentive payments by selecting a small number of participants while still satisfying probabilistic coverage constraint. Although the studies [8, 9] considered the spatio-temporal coverage requirements, they both assumed that only one sensing task was involved. But, many sub-tasks are generally involved in a MCS task, and they may be published on the sensing platform at the same time. Hence, new participant selection mechanisms are needed in order to choose appropriate participants for different sub-tasks.

The main contributions of our work can be concluded as follows:

- Based on large-scale real dataset of “KaiTianYan”, we design a data quality measurement method and data payment strategy for MCS scenario.
- To evaluate the reliability of participants, we propose a quality-aware participant reputation model by introducing active factor to lay a theoretical foundation.
- We propose a data-quality-aware participant selection mechanism. Based on a multistage decision process, a greedy strategy is used to solve the objective optimization problem.

The remainder of this paper can be organized as follows. Section 2 gives task model and task reward. In Sect. 3, a participant reputation model is designed. In Sect. 4, we propose a data-quality-aware participant selection mechanism. Section 5 gives simulation results. Finally, the conclusion is drawn in Sect. 6.

2 Task Model and Task Reward

2.1 Data Set

Since 2015, IoT technology laboratory of BUPT has launched “KaiTianYan” project, which is a MCS air pollution monitoring activity by recruiting participants to have a “Sky Shot”. In this project, a sensing task can be published through APP and pictures collected can be upload to and processed on a server. The dataset contains a total of

31,601 pieces of pictures. The collection period of this project lasts from May 2015 to January 2016. 13 monitoring regions are involved and more than 60 participants are chosen in this project.

2.2 Task Scenario

Task scenario can be described as follows. An air pollution monitoring task lasts from 8:00 am to 18:00 pm in one day, data requester releases a sub-task every 2 h, such as 8:00, 10:00, 12:00. Each sub-task specifies a to-be-sensed region and participants need to complete sub-task before the start of next sensing period.

Generally, the whole task flow consists of six processes: task distribution, sign up, participant selection by introducing reputation, sensing data upload, reputation update and participant reward.

2.3 Payment Strategy

In MCS, the sensing platform needs to pay participants after receiving their sensing data. Considering that the reference basis (e.g. reputation, receiving capacity) for payment is not directly proportional to the sensing data quality, we propose a hybrid pay strategy by mixing basic reward with dynamic reward for a MCS task. For each participant i , r_i denotes his/her reputation, q_{ij} denotes the quality of the j^{th} sub-task, and ST_RWD_j is the j^{th} sub-task's reward. So, the reward RWD_i can be calculated by Eq. (1).

$$RWD_i = r_i * BR + \sum_{j=1}^n q_{ij} ST_RWD_j \quad s.t. BR > ST_RWD_j \quad (1)$$

where $r_i * BR$ represents basic reward, which can be affected by budget and expected participant scale, and the sum of $q_{ij} * ST_RWD_i$ represents dynamic reward. Basic reward can be paid only if a selected participant completes a certain sub-task. Differently, a participant can get dynamic reward by joining more than one sub-tasks. So, we can see that the more sub-tasks are completed, the more dynamic reward participants can get.

3 Participant Reputation Model

3.1 Reputation Model

Data Quality

Data quality can be measured by combining integrity and accuracy. Integrity refers to the sky part in a picture should make up more than $2/3$. A piece of sensing data (picture) without meeting this above requirement will be marked as *junk*. Otherwise,

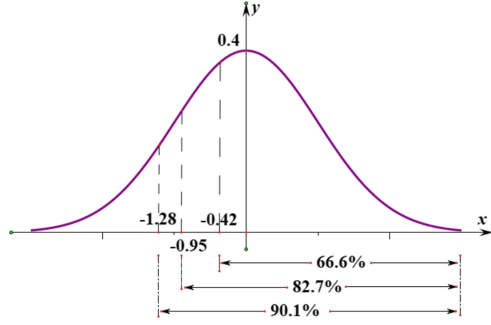


Fig. 1. The mapping relationship of data quality

we qualify data quality according to Air Quality Index (AQI) in China. $AQI \in [0, 50]$ can be defined as excellent; in turn, $AQI \in [51, 100]$ is good, $AQI \in [101, 150]$ is lightly polluted. We define the pollution index difference between calculated value from “sky shot” picture uploaded and the monitoring value from monitoring station as Δq . If $\Delta q \leq 50$, then the uploaded picture can be labeled as *high* quality; similarly, if $50 < \Delta q \leq 100$ or $\Delta q > 100$, the data quality can be labeled as *middle* quality or *low* quality, respectively.

Through the statistical analysis on the real dataset, we find out the data quality satisfies a normal distribution. As illustrated in Fig. 1, *high* quality reaches 66.6%, the *medium* and *low* quality data accounts for 23.5%, and the *junk* quality data is about 9.9%. By comparing the standard normal distribution table, we give a quantitative method to calculate data quality q_l , as given in Eq. (2). For simplicity, data quality value can be normalized in $[0, 1]$.

$$q_l = \frac{1}{6} (\Phi^{-1}(\sum_{l=junk}^{level} Pct_l) + 3), \quad (2)$$

$$level = \{high, middle, low, junk\}$$

Willingness

Willingness represents participants’ social attributes. Assuming the sensing platform releases sub-tasks at discrete time t_s and finishes at t_e ($t_e - t_s = 2$ h). Besides, participant i is assumed to complete a sub-task at discrete time t . If a participant uploads sensing data within 1 h after t_s , which means the participant has high participation wish. Otherwise, willingness will have a great attenuation. For participant i and the j^{th} sub-task, its willingness w_{ij} can be calculated by Eq. (3), and it is quantified within $[0, 1]$.

$$w_{ij} = \frac{-\alpha \text{Arctan}(\beta(\text{time} - (t_s - t/2)))}{0.5\pi} + l; \quad t_s \leq \text{time} \leq t_e \quad (3)$$

In conclusion, data quality is an objective factor and willingness is a subjective factor and. For participant i and the j^{th} sub-task, its corresponding reputation r_{ij} can be measured by mapping f to \ln function, as defined in Eq. (4). f is the fusion result of data quality and willingness $f(q_{ij}, w_{ij}) = a_j q_{ij} + b_j w_{ij}$, where a_j, b_j are respectively weight coefficients.

$$r_{ij} = \begin{cases} \ln(f) & , f > 1 \\ -\ln(-(f - 2)) & , 0 \leq f \leq 1 \end{cases} \quad (4)$$

3.2 Reputation Update

For the n^{th} (time) sub-task for participant i , its historical reputation can be defined as r_i^n . Specially, the sensing platform sets r_i^0 as 0.5 for each new participant. Here, a data fusion based logistic regression method is employed for reputation update. As described in Eq. (5), where $\frac{10(r_i^n + r_{ij})}{\max(r_i^n + r_{ij}) - \min(r_i^n + r_{ij})} - 5$ denotes definition domain transformation which ranges from $[0, 1]$ to $[-5, 5]$, and Rep_i denotes the updated reputation.

$$Rep_i = 1 / (1 + e^{-\frac{10(r_i^n + r_{ij})}{\max(r_i^n + r_{ij}) - \min(r_i^n + r_{ij})} - 5}) \quad (5)$$

Reputation update with logistic regression function has some significant advantages. For example, it can magnify the impact of the current reputation to the overall reputation and thus motivate participants to collect high-quality data. However, when the overall reputation is close to 1, the impact of the current reputation becomes smaller and tends to be stable. It means that this reputation model cannot effectively distinguish active participant and inactive one. Here, this problem can be resolved by introducing activity factor. For participant i , his/her activity factor A_i can be defined as Eq. (6), where K is the number of sub-tasks completed by participant i , P is a threshold (the average number of sub-tasks completed by all participants), and α is the number of submission by the most active participant.

$$A_i = \frac{\text{Arctan}(K - P) + \text{Arctan}P}{\pi + 2\text{Arctan}P} + 0.5, P \in [0, \alpha] \quad (6)$$

Therefore, a participant reputation can be updated by $Rep'_i = Rep * A_i$.

4 Participant Selection Mechanism

In this section, we propose a data-quality-aware participant selection mechanism with satisfying multi-objective optimization.

4.1 Multi-objective Optimization

Here, we realize the multi-objective participant selection optimization by using constraint method. We translate the pending problem into a single-objective one which is easier to solve. Specifically, the main objective $f_k(x)$ can be determined, other $k - 1$ objectives are considered as constraint conditions, which can be defined as follows.

$$\begin{aligned} & \max f_k(x) \\ & s. t. f_i(x) \in \tau_i (i = 1, 2, \dots, k - 1) \end{aligned}$$

In the participant selection process, the focus is to maximize $f_k(x)$ while meeting other constraint conditions.

In our work, the following two objectives need to be minimized:

- To minimize cost, the sensing platform chooses participants as few as possible to make BR ($BR > STR$) minimal.
- To minimize participant scale, each participant needs to complete at least q sensing sub-tasks in a full task cycle which lasts T hours. Each sub-tasks lasts t hours.

The following two constraint conditions are also satisfied:

- To ensure data quality, a participant's reputation need reach the threshold and a same device cannot be allowed to upload data repeatedly.
- To achieve task coverage, each sub-task should be covered by at least K participants.

S_{ix} represents the sum of sub-tasks performed by participant i in T/t stages. The objective function and constraint conditions are defined as follows:

$$\left\{ \begin{array}{l} \min \sum_{i=1}^n \text{Cost}(p_i) \\ \min(P.\text{scale}), \sum_{x=1}^{T/t} S_{ix} \geq q \end{array} \right. \quad s.t. \quad \left\{ \begin{array}{l} \max(\sum_{i=j=1}^{i=n, j=m} q_{ij}) \\ \max(\text{Covr}), |U_j| = K (1 \leq j \leq m) \end{array} \right.$$

Participant Selection

A continuous MCS task can be divided into several stages from time domain, and each stage is an independent problem. Considering that the selection result for each stage will affect the next stage, that is, $P^{(x+1)} = \Psi_x(F; P^{(0)}, P^{(1)}, \dots, P^{(x)})$, we adopt greedy strategy for each stage.

The set of sub-tasks in a sensing task can be defined as $\text{Task} = \{task_1, task_2, \dots, task_m\}$, x ($x = 1, 2, \dots, T/t$) denotes a certain stage, the user set in one stage can be denoted as $U = \{u_1, u_2, \dots, u_n\}$, the participant set who performs $task_j$ can be denoted as $U_j = \{u_{j1}, u_{j2}, \dots\}$. A Multi-Stage Decision method based on Greedy strategy (MSD-G) is proposed to solve the problem (see Algorithm 1).

Algorithm 1 MSD-G Algorithm

Input: User set $U = \{u_1, u_2, \dots, u_n\}$
 For each user $u_i \in U$, he/she has:
 Num of sub-task completed $S_{ix} = \{l_{i1}, l_{i2}, \dots, l_{in}\}$
 Region set $L = \{l_1, l_2, \dots, l_w\}$

Initial: Candidate set $C_x = \emptyset$; Participant Set: P ; K -coverage
 Coverage array: $H_{x \times w} = \{[0, 0, \dots, 0], \dots\}$

```

1: for all  $u_i \in U$ 
2:   do  $A(u_i) \leftarrow \text{getATP}_i(\text{activity\_factor}, \text{reputation})$ ;
3:   if  $A(u_i) \geq \text{threshold}$  then
4:     do  $c_i \leftarrow u_i$ ;
5:   end if
6: end for
7: while  $L \neq \emptyset$ 
8:    $p_i \leftarrow \text{findBest}(C_x, P_{x-1})$ ;
9:    $C_x \leftarrow C_x - \{c_i\}$ ;
10:   $P_x \leftarrow P_x \cup \{p_i\}$ ;
11:  for all  $S_{i1}$  do
12:    if  $H[x, h] \leq K$  then
13:      do  $H[x, h] \leftarrow H[x, h] + 1$ ;
14:    else
15:       $L \leftarrow L - \{l_h\}$ ;
16:    end if-else
17:  end for
18: end while
19: for  $i \leftarrow 1$  to  $P_x.\text{Scale}$  do
20:    $q_{ij} \leftarrow$  calculate data quality level
21:    $w_{ij} \leftarrow$  calculate willingness value
22:    $e_{ij} \leftarrow$  data fusion
23:    $r_i^n \leftarrow$  calculate reputation state
24:    $\text{Rep}_i \leftarrow$  update reputation
25:    $\text{ATP}_i \leftarrow$  update  $\text{ATP}$ 
26: end for
27: return  $P_x$ 

```

5 Simulation Results

We set $\chi = 1$, $\lambda = 0.5$. According to Eq. (3), the ST_RWDs of multiple sub-tasks can be calculated and shown in the form of contour, which can provide references for cost budget. As illustrated in Fig. 2, contour infers that regions with darker color and slower slope have excellent task completion. In the regions where users' resources are rich, the sensing platform can appropriately reduce their corresponding budgets, and thus make up for the regions where users' resources are poor.

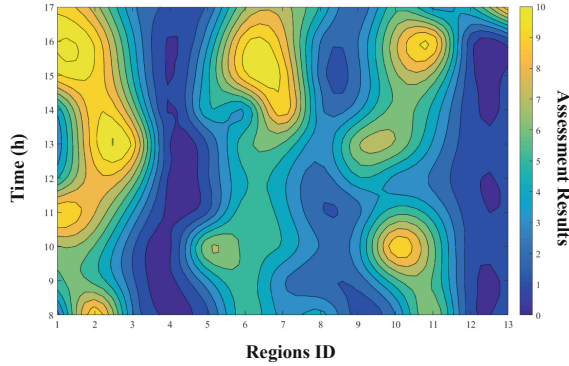


Fig. 2. Contour map of ST_RWDs of multiple sub-tasks

Reputation Model Assessment

To evaluate the performance of the proposed solution, data set can be divided into two parts. The participant reputation without and with activity factor can be given in Fig. 3. Rich and poor user resource region are respectively represented by \circ and \times . We can find out the reputation model with activity factor can effectively distinguish between high active participants and low ones.

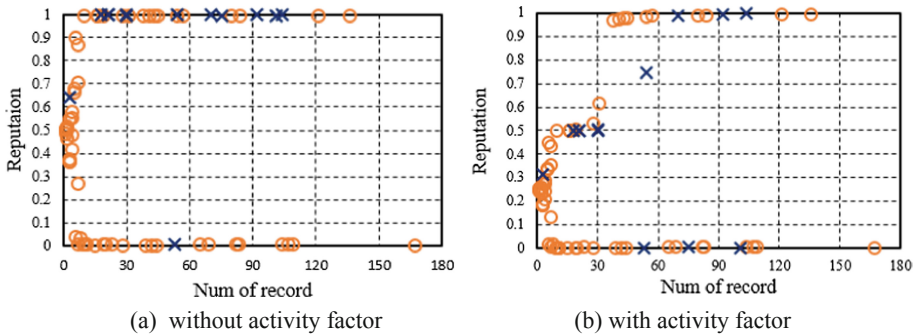


Fig. 3. Comparison of participant reputation

5.1 Performance Analysis

In this section, two popular participant selection mechanisms: Random Sort and First Come First Served mechanism (FCFS) are compared with our proposed MSD-G. Simulation parameter setting is listed in Table 1.

Table 1. Simulation parameter setting

Parameter	Value (default)
Task duration (T)	10 h
Sub-task duration (t)	2 h
Number of Regions	3
Reputation threshold (ATP)	0.5
K-coverage	3-coverage

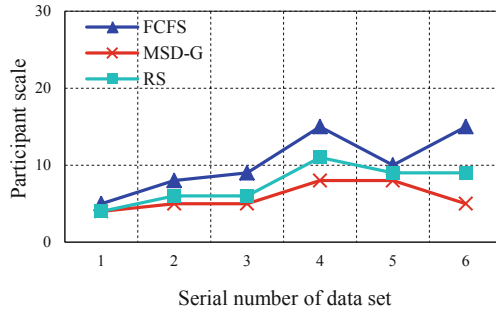
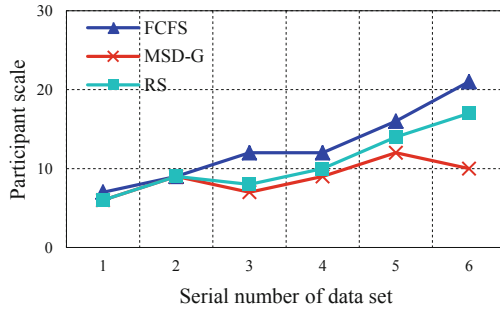
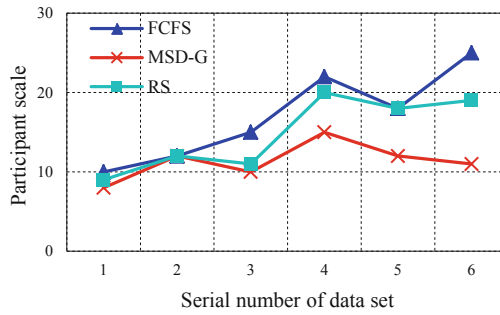
Analysis of Different K

In air pollution monitoring application, scenario can be described as follows: In each stage, users sign up for sensing sub-task at first. Then users with good reputation have chances to be chosen to carry on sensing sub-task during 8:00 am to 18:00 pm, and each region is covered by K times.

The experimental parameters are set as $T = 10$, $t = 2$ and $ATP \geq 0.5$. As shown in Fig. 4, we can find that with the increase of K , the participant scales for three kinds of participant selection mechanisms increase correspondingly. FCFS has the biggest participant scale, mainly because that the selected participants have the strongest participation will. However, its efficiency is the best. Moreover, with the increase of problem complexity, the participant scale of MSD-G decreases significantly and keeps stable relatively. In terms of resource utilization, MSD-G only needs a small amount of participants to complete the same sensing task. In all, our MSD-G mechanism can complete a continuous monitoring task with the lowest (personnel) cost.

Analysis of Different t

The experimental parameters are set as $T = 10$, $K = 3$ and $ATP \geq 0.5$. As seen in Fig. 5, with the decrease of the stage duration, the participant scale will increase. When the sensing frequency is low, there is no big difference between three selection mechanisms. With the increase of sensing frequent, the performance of FCFS on resource utilization becomes worse and Random Sort gets more unstable while MSD-G achieves the best performance.

(a) $K=2$ (b) $K=3$ (c) $K=4$ **Fig. 4.** Comparison of participant scale with different K

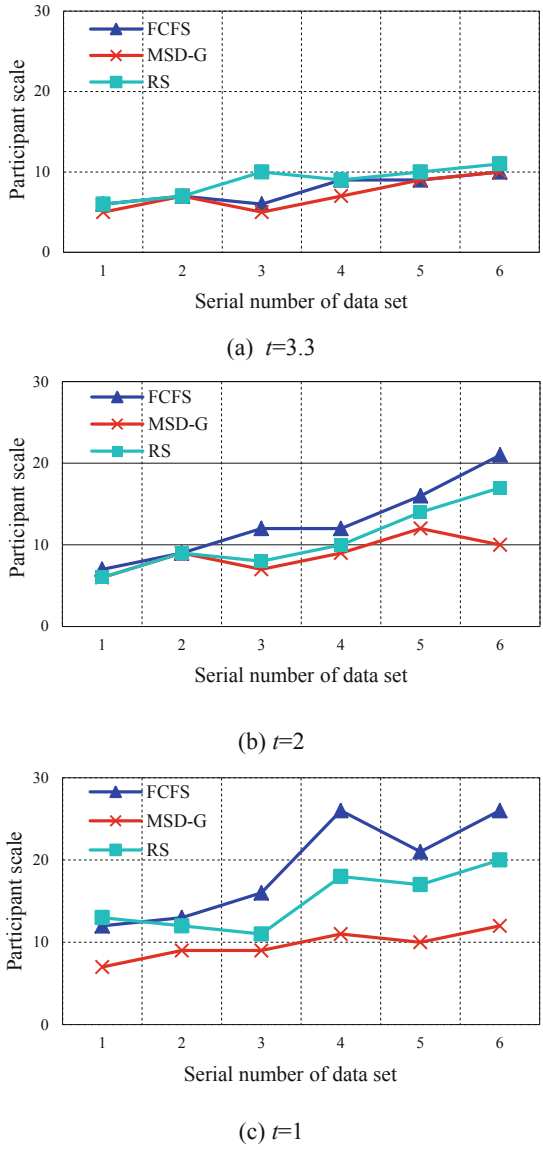


Fig. 5. Comparison of participant scale with different t

6 Conclusion

To assure data quality, in this paper, we propose a quality-aware reputation model by introducing activity factor. Particularly, we present a data-quality-aware participant selection mechanism for MCS system. Simulation results show that MSD-G can accomplish sensing tasks with optimizing the data quality, cost and participant scale.

Acknowledgment. This work was supported by the National Natural Science Foundation of China under Grant No. 61872027. Thanks Prof. Liang Liu in IoT technology laboratory of Beijing University of Posts and Telecommunications for providing all the materials of “Kai-TianYan” project.

References

1. Guo, B., Zhai, S.Y., Yu, Z.Y.: Crowdsensing big data: sensing data selection, and understanding. *Big Data Res.* **3**(5), 57–69 (2017)
2. Hu, J., Tao, D.: Theories and methods of quality measure and assurance for mobile crowd sensing. *J. Chin. Comput. Syst.* **40**(5), 918–923 (2019)
3. Zhao, D., Ma, H.D.: Quality measuring and assurance for mobile crowd sensing. *ZTE Technol. J.* **21**(6), 2–5 (2015)
4. Zhang, X.L., Yang, Z., Sun, W., Liu, Y.H., et al.: Incentives for mobile crowd sensing: a survey. *IEEE Commun. Surv. Tutorials* **18**(1), 54–67 (2016)
5. Tao, D., Zhong, S., Luo, H.: Staged incentive and punishment mechanism for mobile crowd sensing. *MDPI Sens.* **18**(7), 1–21 (2018)
6. Yang, J., Li, P., Wang, H.: Participant reputation aware data collecting mechanism for mobile crowd sensing. In: 2017 IEEE/CIC International Conference on Communications in China (ICCC), Qingdao, pp. 1–6 (2017)
7. Pournajaf, L., Xiong, L., Sunderam, V., Goryczka, S.: Spatial task assignment for crowd sensing with cloaked locations. In: 15th IEEE International Conference on Mobile Data Management, Brisbane, pp. 73–82 (2014)
8. Liu, C.H., Zhang, B., Su, X.: Energy-aware participant selection for smartphone-enabled mobile crowd sensing. *IEEE Syst. J.* **11**(3), 1435–1446 (2017)
9. Zhang, D.Q., Xiong, H.Y., Wang, L.Y.: CrowdRecruiter: selecting participants for piggyback crowdsensing under probabilistic coverage constraint. In: ACM International Joint Conference on Pervasive & Ubiquitous Computing. ACM, Washington (2014)



Infrared Small Target Detection Based on Facet-Kernel Filtering Local Contrast Measure

Peng Du  and Askar Hamdulla 

Xinjiang University, Shengli Road No. 666, Urumqi, Xinjiang, China
askar@xju.edu.cn

Abstract. How to detect small targets accurately under complex background and low signal-to-clutter ratio is of great significance to the development of precision guided weapons and infrared early warning. The traditional local contrast method is difficult to detect small and dim targets in complex background. In this paper, in order to improve the traditional local contrast method and detect small targets effectively under complex background conditions, a novel method base on Facet-kernel filtering local contrast measure (FFLCM) is proposed for small target detection. Initially, a nest sliding window structure of the central layer and the surrounding background layer is given. Then, the Facet-kernel filter is used to enhance the target in the center layer, the gray similarity difference between the central layer and the surrounding layer is calculated to suppress the background. Finally, a threshold operation is used to extract target. Experimental results demonstrate that our proposed method could effectively enhance small targets and suppress complex background clutters simultaneously.

Keywords: Facet-kernel · Infrared image · Gray similarity difference · Local contrast measure

1 Introduction

Infrared search and track (IRST) systems is widely applied in the various fields, such as precise guidance, pre-warning, remote sensing, aerospace, etc. [1, 2]. Detecting a small IR target of unknown position and velocity at low signal-to-noise ratio (SNR) is an important issue in IR search and track system, which is necessary for military applications to warn from incoming small targets from a distance, such as enemy aircraft and helicopters [3, 4]. It is usually very difficult to detect IR small target because the target is only a dim and small spot [5], it can be easily drowned by complex backgrounds. Moreover, random electrical noise of the detector may cause some pixel-sized noises with high brightness (PNHB) in the image, they will easily be mistaken as targets [6].

In recent years, the local contrast mechanism of human visual system (HVS) [7] is introduced to the field of IR small target detection, for example, Wang et al. [8] proposed a easier filter template named difference of Gaussian (DoG), Although this method can strengthen the target, it can not suppress the background well, resulting in a higher false

alarm rate; Chen et al. [9] proposed the local contrast measure (LCM), it used a nested structure in which the surrounding area is divided into eight directions, Wei et al. [10] proposed the Multi-scale patch-based contrast measure (MPCM), it merged two corresponding directions together; Han et al. [11] proposed a multi-scale relative local contrast measure (RLCM) using both ratio and difference operations; Nie et al. [12] proposed a multi-scale local homogeneity measure (MLHM), it considered the homogeneity of the central area, Wang et al. [14] proposed a Facet detection method; Qin et al. [15] proposed the Novel Local Contrast Measure (NLCM) and so on.

Generally speaking, LCM and the improved algorithm can suppress the background by using the ratio or difference relationship between the central domain and the surrounding domain, but often weaken the brightness and shape of the real target. Especially in complex background, it is difficult to detect the small real target after the brightness is weakened and the shape is reduced. In addition, the common LCM algorithm uses multi-scale computation, which will result in too long calculation time, meanwhile, different scales require different parameters is a very troublesome work.

In this paper, an effective IR small target detection based on facet-kernel local contrast measure is presented. First, we give a sliding window with new nest structure which include central layer and surrounding background layer. Then, in the central layer, the facet kernel model is calculated to strengthen the target. In the surrounding layer, the average gray value of each surrounding cell is calculated to suppress background. Finally, we use a threshold operation to extract the target.

2 The Proposed Method

2.1 Construction of a New Nest Structure

In this letter, a nest structure [9] with two layers is proposed, as shown in Fig. 1. The central layer with size $c \times c$ is used to capture a target, it doesn't need to adjust its size to the target size and can deal with targets of different sizes by a simple single-scale calculation, so the computations can be reduced significantly.

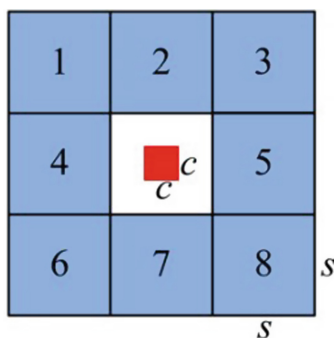


Fig. 1. A nest structure used in the proposed algorithm. The red square is the facet-kernel model size window, the white square is the central layer, blue squares numbered with 1–8 are the surrounding layer with eight directions. (Color figure online)

The surrounding layer is used to capture the surrounding background of the target, the surrounding layer is divided into eight directions (cell(1)–cell(8)) in Fig. 1. The cell size is $s \times s$. According to the definition of SPIE, a small target shouldn't be larger than 9×9 , so we set s to 9 here. The cell size s of the surrounding background layer is set to 9×9 . The central layer design is slightly larger than the target because the gray gradient should be taken into account when the facet-kernel filter be calculated.

2.2 FFLCM Calculation

In each images, slide the sliding window shown in Fig. 1 from top to bottom and left to right. In each sliding window, we calculate the FFLCM. Finally, the saliency map (FFLCM map) of the test results will be obtained. The calculation process will be described in detail below.

Enhancing the Target in the Central Layer

A high frequency filter operator, facet kernel filter is used. Its advantage is that it has fast processing speed, which can enhance the different sizes of targets quickly and suppress the background clutter. The design idea of the facet kernel model is that the facet model function is used to estimate the gray surface of each pixel area in the pre-processed image. It can represent the gradient magnitude information in different directions of the image. The facet kernel convolution can enhance the target area, and is not easily disturbed by image noise and clutter.

Specifically, for input central-layer images $I_c(x, y)$, the formula of facet kernel filtering is:

$$I_f(x, y) = I_c(x, y) * f_{w5 \times 5}. \tag{1}$$

$f_{w5 \times 5}$ is a facet-kernel, $I_f(x, y)$ is the result of central layer filtering,* for convolution operations.

The facet kernel model is defined as a small neighborhood in the central layer, here, we adopt size is 5×5 and the intensity surface of discrete pixels can be approximated by a binary cubic polynomial.

First a symmetric set R is defined as $R = \{-2, -1, 0, 1, 2\}$ and the discrete orthogonal polynomial set is: $\{1, r, r^2 - 2, r^3 - 17/5r, r^4 + 3r^2 + 72/35\}$, another symmetric set C is determined similarly as $C = \{-2, -1, 0, 1, 2\}$, with its discrete orthogonal polynomial set: $\{1, c, c^2 - 2, c^3 - 17/5c, c^4 + 3c^2 + 72/35\}$, Afterward, ten 2-D discrete orthogonal Chebyshev polynomials π_i ($i = 1, 2, \dots, 10$) are constructed by ignoring the orders higher than 3.

$$\pi_i \in \{1, r, c, r^2 - 2, rc, c^2 - 2, r^3 - (17/5)r, (r^2 - 2)c, r(c^2 - 2), c^3 - (17/5)c\}. \tag{2}$$

Finally, the pixel surface function $f(r, c)$ in this $R \times C$ area is given in (3) and the coefficients k_i ($i = 1, 2, \dots, 10$) could be deduced by the least-squares algorithm.

$$f(r, c) = \sum_{i=1}^{10} kipi. \tag{3}$$

Based on the orthogonal property of polynomials, we can simplify the calculation of k_i as shown in (4), where $I(r, c)$ represents the original pixel values in $R \times C$ area.

$$k_i = \frac{\sum_{(r,c) \in R \times C} pi(r, c) I(r, c)}{\sum_{(r,c) \in R \times C} pi^2(r, c)}. \tag{4}$$

Furthermore, $I(r, c)$ is independent of the remainder in (4), which could be viewed as a fixed filter denoted by w_i in (5). Thus, k_i could be calculated directly through convolution using the corresponding w_i .

$$w_i = \frac{pi(r, c)}{\sum_{(r,c) \in R \times C} pi^2(r, c)}. \tag{5}$$

Therefore, according to formulas (2) and (3), the second-order partial derivatives of window central pixels (0, 0) along row (0°) and column (90°) directions can be obtained:

$$\frac{\partial^2 f(r, c)}{\partial r^2} = 2K4; \quad \frac{\partial^2 f(r, c)}{\partial c^2} = 2K6. \tag{6}$$

It can be seen that the target can be enhanced by calculating the sum of coefficients $K4$ and $K6$. $K4$ and $K6$ can be calculated by formula (5) and convoluted with $I(x, y)$. Substitute $pi(r, c)$ into formula (5),

$$w4 = \frac{1}{70} \begin{bmatrix} 2 & 2 & 2 & 2 & 2 \\ -1 & -1 & -1 & -1 & -1 \\ -2 & -2 & -2 & -2 & -2 \\ -1 & -1 & -1 & -1 & -1 \\ 2 & 2 & 2 & 2 & 2 \end{bmatrix}, \quad w6 = w4^T. \tag{7}$$

Therefore, $f_{w5 \times 5} = -2(W4 + W6)$, the inverse is used to detect bright targets whose center gray level is greater than edge gray level.

The facet-kernel filter is finally expressed as:

$$f_{w5 \times 5} = \begin{bmatrix} -4 & -1 & 0 & -1 & -4 \\ -1 & 2 & 3 & 2 & -1 \\ 0 & 3 & 4 & 3 & 0 \\ -1 & 2 & 3 & 2 & -1 \\ -4 & -1 & 0 & -1 & -4 \end{bmatrix}. \tag{8}$$

After filtering with this facet-kernel filter in the central layer, the target will be significantly enhanced.

Suppressing the Background Clutter Between Central and Surrounding Layer

It can be seen that the target is enhanced by using facet kernel filtering in the central layer, and the false alarm rate needs to be further reduced by suppressing background clutter. The background layer can be suppressed by calculating the similarity difference between the central layer and the surrounding layer blocks in eight directions.

$$M_{isur} = \frac{1}{n} \sum_{j=1}^n I_j^i, (i = 1, 2, \dots, 8),$$

$$d_i = (\max I_f(x, y) - M_{isur})^2, (i = 1, 2, \dots, 8). \quad (9)$$

where n is the number of the pixels in the i th cell and I_j^i is the gray level of the j th pixel in the i th cell, M_{isur} denotes the gray mean of eight directions of the surrounding layer, d_i represents similarity difference, $\max I_f(x, y)$ represents the maximum gray value of the pixels in the $I_f(x, y)$.

FFLCM Calculation

$$FFLCM(x, y) = \frac{meandi - mindi}{maxdi - mindi}, (i = 1, 2, \dots, 8). \quad (10)$$

where $meandi$, $maxdi$ and $mindi$ represent the gray mean, maximum and minimum of d_i , respectively.

2.3 Threshold Operation

For each pixel of the raw image, construct a nest sliding window and calculate the corresponding FFLCM according to (1), (9) and (10), then form the results as a new matrix named saliency map(SM). A simple threshold operation [13] will be used to extract the true target, and the threshold is defined as

$$Th = \lambda \max_{SM} + (1 - \lambda) mean_{SM}. \quad (11)$$

where \max_{SM} and $mean_{SM}$ are the maximum and average of SM, respectively. λ is a given factor, our experiments show that 0.5–0.7 will be proper for single target detection.

3 Experimental Results

Based on the traditional LCM algorithm, the FFLCM method in this paper is improved. In order to fully demonstrate the advantages of the proposed algorithm, the three-dimensional gray distribution map of the image processed by the algorithm is given intuitively, as shown in Fig. 2. In addition, two indicators, SCRG (signal to clutter ratio gain) and BSF (background suppression factor), are used to evaluate the target

enhancement ability and background suppression ability of the algorithm. The definitions of the two indicators are shown in formula (12):

$$SCRG = \frac{SCR_{out}}{SCR_{in}}, BSF = \frac{\sigma_{in}}{\sigma_{out}}. \tag{12}$$

where SCR_{in} and SCR_{out} are the SCR values of the raw image and SM map, respectively, and σ_{in} and σ_{out} are the standard deviation of the raw image and SM map, respectively.

Table 1. SCRG values for the different algorithms

	DoG	LCM	MPCM	MLHM	RLCM	FACET	NLCM	Proposed
a(1)	4.0830	10.5809	3.2303	6.3453	11.2212	5.0528	10.2832	30.4508
c(1)	4.6455	12.2334	4.0375	7.5058	14.2337	6.0223	12.3543	32.4646

Table 2. BSF values of the different algorithms

	DoG	LCM	MPCM	MLHM	RLCM	FACET	NLCM	Proposed
a(1)	2.3433	1.3467	1.5923	4.0972	5.7634	3.1022	4.9812	20.0818
c(1)	3.4532	2.0286	2.8329	5.1159	8.2632	3.2893	6.2321	30.0105

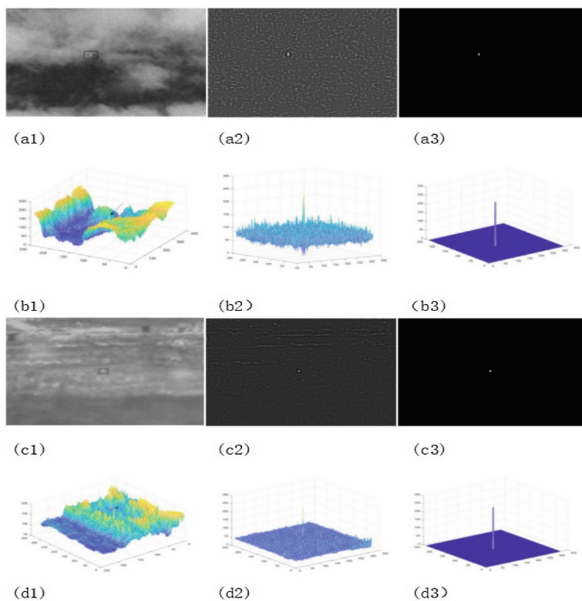


Fig. 2. a(1), c(1) are two sets of original maps, a(2), c(2) are SM maps, a(3), c(3) are result maps after threshold operation. (b1–b3), (d1–d3) are (a1–a3), (c1–c3) corresponding to the three-dimensional gray distribution map.

From the Fig. 2, we can see that our proposed method can correctly extract small and dim targets in complex background environment, and there is almost no background clutters.

In Tables 1 and 2, we compare other similar methods, from which we can see that our proposed methods have improved the ability of target enhancement and background suppression. Experiments show that our proposed method is more effective than the other compared methods.

4 Conclusion

In this letter, a new Facet-kernel filtering local contrast measure (FFLCM) for IR small target detection is proposed, it can deal with different sizes of targets using only single-scale calculation, and can enhance true target and suppress complex backgrounds simultaneously. Experimental results show that our proposed FFLCM algorithm can achieve a good detection performance.

Acknowledgments. This work has been supported by the National Natural Science Foundation of China (No. 61563049).

References

1. Deng, H., Sun, X., Liu, M., Ye, C., Zhou, X.: Small infrared target detection based on weighted local difference measure. *IEEE Trans. Geosci. Remote Sens.* **54**(7), 4204–4214 (2016)
2. Nasiri, M., Mosavi, M.R., Mirzakuchaki, S.: Infrared dim small target detection with high reliability using saliency map fusion. *IET Image Process.* **10**(7), 524–533 (2016)
3. Kim, S., Lee, J.: Scale invariant small target detection by optimizing signal-to-clutter ratio in heterogeneous background for infrared search and track. *Pattern Recogn.* **45**(1), 393–406 (2012)
4. Zhang, L.F., Zhang, L.P., Tao, D., Huang, X.: A multifeature tensor for remote-sensing target recognition. *IEEE Geosci. Remote Sens. Lett.* **8**(2), 374–378 (2011)
5. Bai, X., Bi, Y.: Derivative entropy-based contrast measure for infrared small-target detection. *IEEE Trans. Geosci. Remote Sens.* **56**(4), 2452–2466 (2018)
6. Han, J., Ma, Y., Zhou, B., Fan, F., Liang, K., Fang, Y.: A robust infrared small target detection algorithm based on human visual system. *IEEE Geosci. Remote Sens. Lett.* **11**(12), 2168–2172 (2014)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11), 1254–1259 (1998)
8. Wang, X., Lv, G., Xu, L.: Infrared dim target detection based on visual attention. *Infrared Phys. Technol.* **55**(6), 513–521 (2012)
9. Chen, C.P.L., Li, H., Wei, Y., Xia, T., Tang, Y.Y.: A local contrast method for small infrared target detection. *IEEE Trans. Geosci. Remote Sens.* **52**(1), 574–581 (2014)
10. Wei, Y., You, X., Li, H.: Multiscale patch-based contrast measure for small infrared target detection. *Pattern Recogn.* **58**, 216–226 (2016)

11. Han, J., Liang, K., Zhou, B., Zhu, X., Zhao, J., Zhao, L.: Infrared small target detection utilizing the multi-scale relative local contrast measure. *IEEE Geosci. Remote Sens. Lett.* **15**(4), 612–616 (2018)
12. Nie, J., Qu, S., Wei, Y., Zhang, L., Deng, L.: An infrared small target detection method based on multiscale local homogeneity measure. *Infrared Phys. Technol.* **90**, 186–194 (2018)
13. Fu, H., Long, Y., Zhu, R., An, W.: Infrared small target detection based on multiscale center-surround contrast measure. In: *Proceedings of SPIE*, vol. 10615, pp. 106150I–106150I-08 (April 2018)
14. Wang, G.-D., Chen, C.-Y., Shen, X.-B.: Facet-based infrared small target detection method. *Electron. Lett.* **41**(22), 1244–1246 (2005)
15. Qin, Y., Li, B.: Effective infrared small target detection utilizing a novel local contrast method. *IEEE Geosci. Remote Sens. Lett.* **13**(12), 1890–1894 (2016)

Author Index

- Cai, Qingsong 3
Cao, Yuan 317
Chen, Chao 289
Chen, Guihai 61
Chen, Junji 143
- Dan, Jingpei 143
Dang, Xiaochao 169, 302, 317, 332
Du, Peng 360
Duan, Yongshuai 61
- Feng, Guangsheng 35, 263
- Gao, Rong 275
Gu, Peiyuan 127
Guo, Longjiang 224, 236
- Hamdulla, Askar 360
Hao, Zhanjun 169, 302, 317, 332
He, Di 35
He, Yunhua 198
Huang, Xia 143
- Jie, Huilin 289
Jin, Feiyu 289
- Kong, Linghe 61
- Li, Bingyang 35
Li, Caixia 317
Li, Guiduan 212
Li, Jitong 198
Li, Ming 112
Li, Peng 236
Li, Quanming 35, 263
Li, Wen 212
Li, Xinyue 251
Li, Yaoping 251
Li, Yong 101
Li, Yuexia 169, 332
Lin, Jia 3
Lin, Junyu 263
Liu, Hong 236
- Liu, Kai 289
Liu, Ling 143
Liu, Wei 48
Liu, Xingcheng 90
Lu, Anqi 154
Luo, Rong 48
Lv, Hongwu 35, 263
Lv, Silin 263
- Meng, Fanrong 112
Meng, Tong 61
- Nie, Wenmei 101
- Pang, Chengjie 198
Peng, Jun 90
- Ren, Meirei 224
- Song, Guozhi 212
Song, Tian 127
Song, Xiaoxia 101
Sun, Hongbin 348
Sun, Limin 198
- Tan, Li 19
Tang, Xiaojiang 19
Tao, Dan 348
Tian, Zengshan 251
- Wang, Ana 224
Wang, Chao 198
Wang, Haoyu 19
Wang, Huiqiang 35, 263
Wang, Jing 76
Wang, Liangmin 184
Wang, Xiaoming 236
Wang, Yuming 143
Wang, Zhe 61
Wang, Zi-hao 76
Wu, Fan 61
Wu, Hejun 275
Wu, Xiaojun 236

Xia, Yu 48

Xiang, Chaocan 289

Xiong, Shuming 184

Xu, Yabin 127

Yan, Bin 224

Yan, Lihua 302

Yang, Minghua 19

Yang, Zhongheng 275

Yu, Yue 112

Yuan, Hui 251

Zhang, Hao 289

Zhang, Lichen 224, 236

Zhang, Mingzheng 184

Zhang, Tong 169, 332

Zhou, Mu 251

Zhu, Jinghua 154