Smart Innovation, Systems and Technologies 128

Yong Zhao
Tsu-Yang Wu
Tang-Hsien Chang
Jeng-Shyang Pan
Lakhmi C. Jain *Editors*

Advances in Smart Vehicular Technology, Transportation, Communication and Applications

Proceeding of the Second International Conference on Smart Vehicular Technology, Transportation, Communication and Applications, October 25–28, 2018, Mount Emei, China, Part 2

KES International

Springer

Smart Innovation, Systems and Technologies

Volume 128

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

More information about this series at http://www.springer.com/series/8767

Yong Zhao · Tsu-Yang Wu
Tang-Hsien Chang · Jeng-Shyang Pan
Lakhmi C. Jain
Editors

Advances in Smart Vehicular Technology, Transportation, Communication and Applications

Proceeding of the Second International Conference on Smart Vehicular Technology, Transportation, Communication and Applications, October 25–28, 2018, Mount Emei, China, Part 2

Springer

Editors
Yong Zhao
College of Physics and Energy
Fujian Normal University
Fuzhou, Fujian, China

Tsu-Yang Wu
College of Information Science
and Engineering
Fujian University of Technology
Fuzhou, Fujian, China

Tang-Hsien Chang
School of Transportation
Fujian University of Technology
Fuzhou, Fujian, China

Jeng-Shyang Pan
College of Information Science
and Engineering
Fujian University of Technology
Fuzhou, Fujian, China

Lakhmi C. Jain
University of Technology Sydney
Sydney, NSW, Australia

ISSN 2190-3018 ISSN 2190-3026 (electronic)
Smart Innovation, Systems and Technologies
ISBN 978-3-030-04584-5 ISBN 978-3-030-04585-2 (eBook)
https://doi.org/10.1007/978-3-030-04585-2

Library of Congress Control Number: 2018961590

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This volume composes the proceedings of Second International Conference on Smart Vehicular Technology, Transportation, Communication and Applications (VTCA 2018), which is hosted by Fujian University of Technology and is held in Mount Emei, Sichuan Province, China, on October 25–28, 2018. VTCA 2018 is technically co-sponsored by Springer, Southwest Jiaotong University, Fujian University of Technology, Chang'an University, Shandong University of Science and Technology, Fujian Provincial Key Lab of Big Data Mining and Applications, and National Demonstration Center for Experimental Electronic Information and Electrical Technology Education (Fujian University of Technology). It aims to bring together researchers, engineers, and policymakers to discuss the related techniques, to exchange research ideas, and to make friends.

Fifty-six regular papers were accepted in this proceeding. We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, program committee members, and the local committee members for making this conference successful. Finally, we would like to express our special thanks for the financial support from Fujian University of Technology, China, in making VTCA 2018 possible, and also appreciate the great help from Southwest Jiaotong University for locally organizing the conference.

September 2018

Shaoquan Ni
Tsu-Yang Wu
Tang-Hsien Chang
Jeng-Shyang Pan
Lakhmi C. Jain

Organizing Committee

Honorary Chairs

Xinhua Jiang	Fujian University of Technology, China
Lakhmi C. Jain	University of Canberra, Australia; Bournemouth University, UK

Advisory Chair

Xin Tong	Fujian University of Technology, China

Conference Chairs

Shaoquan Ni	Southwest Jiaotong University, China
Yong Zhao	Fujian Normal University, China
Ruimin Li	Tsinghua University, China
Xinguo Jiang	Southwest Jiaotong University, China
Wanjing Ma	Tongji University, China
Yaojan Wu	University of Arizona, USA
Tang-Hsien Chang	Fujian University of Technology, China

Program Chair

Jeng-Shyang Pan	Fujian University of Technology, China

Publication Chairs

Tsu-Yang Wu	Fujian University of Technology, China
Trong-The Nguyen	Hai Phong Private University, Vietnam

Local Organization Chairs

Huoming Shen	Southwest Jiaotong University, China
Zengan Gao	Southwest Jiaotong University, China
Yong Zhang	Southwest Jiaotong University, China

Finance Chair

Jui-Fang Chang	National Kaohsiung University of Science and Technology, Taiwan

Program Committees

Chien-Ming Chen	Harbin Institute of Technology, Shenzhen, China
Dingjun Chen	Southwest Jiaotong University, China
Shu-Chuan Chu	Flinders University, Australia
Cheng Han	Beijing National Railway Research & Design Institute of Signal & Communication Co., Ltd., China
Xingjian Huang	Southwest Jiaotong University, China
Lingming Jiang	Beijing National Railway Research & Design Institute of Signal & Communication Co., Ltd., China
Zongping Li	Jiaotong University, China
Lyuchao Liao	Fujian University of Technology, China
Hunton Lin	Cheng Shiu University, China
Jerry Chun-Wei Lin	Western Norway University of Applied Sciences, Norway
Jun Liu	Beijing National Railway Research & Design Institute of Signal & Communication Co., Ltd., China
Lan Liu	Southwest Jiaotong University, China
Ling Liu	Beijing National Railway Research & Design Institute of Signal & Communication Co., Ltd., China
Ying-Chih Lu	Fujian University of Technology, China
Hongxia Lv	Southwest Jiaotong University, China
Si Ma	Southwest Jiaotong University, China
Jinshan Pan	Southwest Jiaotong University, China
Pei-Wei Tsai	Swinburne University of Technology, Australia
Yuh-Ming Tseng	National Changhua University of Education, Taiwan

Contents

Intelligent Computing Techniques and Its Applications (Invited Session 02)

Emerging Techniques and Its Applications in Computer Science (Invited Session 01)

Modified Adaptive Nonsmooth Attitude Tracking Control of Quadrotor UAV with Dynamic Uncertainties

Dongwei He[1,2](✉), Pei Gao[3], Lisang Liu[1,2], Jing Huang[1,2], Jianxing Li[1,2], and Xuecheng Jiang[4]

[1] Technical Development Base of Industrial Integration Automation of Fujian Province, Fujian University of Technology, Fuzhou 350118, China
yzak_juel@hotmail.com,
{liulisang,huangj,lijx}@fjut.edu.cn
[2] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education, Fujian University of Technology, Fuzhou 350118, China
[3] Department of Information Management Engineering, Fujian Business University, Fuzhou 350012, China
[4] Department of Electronics Engineering, Minjiang University, Fuzhou 350118, China
jxc603@yahoo.com.cn

Abstract. This paper addresses a new attitude controller of quadrotor UAV in the presence of parameters uncertainty and bounded external disturbances. Based on the attitude dynamic model considering parameters uncertainty and bounded external disturbances. Applying backstepping approach, an adaptive nonsmooth attitude controller is developed by integrating nonsmooth control and adaptive control techniques. And the design procedure and stability analysis of the closed-loop system is detailed. Finally, via some simulation results and comparisons, the effectiveness and advantages of the proposed method are illustrated.

Keywords: Quadrotor · Attitude tracking control · Adaptive control
Nonsmooth control

1 Introduction

Quadrotor UAV (Unmanned Aerial Vehicle) has been utilized in a large and expanding number of applications, due to its simple structure, VTOL (Vertical Taskoff and Landing) and great maneuverability [1]. Quadrotor UAV is a typically coupled and underactuated nonlinear system. Who needs attitude tracking to guarantee the quadrotor UAV to track the given flying trajectory. It means that the attitude tracking performance is the basic performance, and high performance tasks like maneuver tasks require high attitude tracking performance. To achieving high attitude tracking performance needs to overcome the coupled character, parameter uncertainties and external disturbance. Apart from the tradition method such as PID, LQ control, robust

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 3–10, 2019.
https://doi.org/10.1007/978-3-030-04585-2_1

feedback linearization, robust control etc., many novel controller methods are proposed. The work [2] proposed a robust disturbance rejection control method, using disturbance observer (DOB) to compensate the uncertainty, similarly, extended state observer was applied to achieve the active disturbance rejection [3, 4]. In [5], a robust control method based on adaptation law was designed to learn and compensate the modeling error and external disturbance. In [6] and [7], model predictive control method was modified and applied to reject the disturbance.

In order to design an attitude tracking controller of quadrotor UAV, which is simple to implement and robust to model uncertainties and external disturbances, in this brief, based on [8], and inspired by [9], an modified adaptive nonsmooth attitude tracking controller with new type structure is developed.

2 Preliminaries

2.1 Model Dynamics of Quadrotor UAV Attitude System

The attitude system model of the rigid-body quadrotor UAV with dynamic uncertainties can be written as (see [1]):

$$\boldsymbol{U} = \boldsymbol{M}\ddot{q} + \boldsymbol{C}(\dot{\boldsymbol{q}}) + \boldsymbol{T} \tag{1}$$

where, $\boldsymbol{q} = [\,\phi \quad \theta \quad \psi\,]^{\mathrm{T}}$, $\dot{\boldsymbol{q}} = [\,\dot{\phi} \quad \dot{\theta} \quad \dot{\psi}\,]^{\mathrm{T}}$, $\ddot{\boldsymbol{q}} = [\,\ddot{\phi} \quad \ddot{\theta} \quad \ddot{\psi}\,]^{\mathrm{T}}$, $\phi \in \mathrm{R}$, $\theta \in \mathrm{R}$, $\psi \in \mathrm{R}$ denote the Euler angles (roll, pitch, yaw) respectively; $\boldsymbol{M} = \mathrm{diag}(\,p_{x1} \quad p_{y1} \quad p_{z2}\,)$, where $p_{x1} = I_x/k_1$, $p_{y1} = I_y/k_1$, $p_{z2} = I_z/k_2$, $k_1 = bl$, $k_2 = d$, $I_x \in \mathrm{R}^+$, $I_y \in \mathrm{R}^+$, $I_z \in \mathrm{R}^+$ are the moments of inertia of UAV, b is the lift coefficient, d is the drag coefficient, l is the distance from the epicenter to the rotor axis; $\boldsymbol{C}(\dot{\boldsymbol{q}}) = [\,-\dot{\theta}\dot{\psi}p_{y1} + \dot{\theta}\dot{\psi}p_{z1} - \dot{\phi}\dot{\psi}p_{z1} + \dot{\phi}\dot{\psi}p_{x1} - \dot{\phi}\dot{\theta}p_{x2} + \dot{\phi}\dot{\theta}p_{y2}]^{\mathrm{T}}$, where $p_{z1} = I_z/k_1$, $p_{x2} = I_x/k_2$, $p_{y2} = I_y/k_2$; $\boldsymbol{T} = [\,T_{\varphi} \quad T_{\theta} \quad T_{\psi}\,]^{\mathrm{T}}$, with: $T_{\phi} = p_{x1}\Delta_{\phi}$, $T_{\theta} = p_{y1}\Delta_{\theta}$, $T_{\psi} = p_{z2}\Delta_{\psi}$, $\Delta_{\phi} \in \mathrm{R}$, $\Delta_{\theta} \in \mathrm{R}$, $\Delta_{\psi} \in \mathrm{R}$ denote the uncertainties of bounded external disturbance; $\boldsymbol{U} = [\,U_{\phi} \quad U_{\theta} \quad U_{\psi}\,]^{\mathrm{T}}$, U_{ϕ}, U_{θ}, U_{ψ} are given as follows:

$$\begin{aligned} U_{\phi} &= \omega_2^2 - \omega_4^2 \\ U_{\theta} &= -\omega_1^2 + \omega_3^2 \\ U_{\psi} &= \omega_1^2 - \omega_2^2 + \omega_3^2 - \omega_4^2 \end{aligned} \tag{2}$$

Where $\omega_i \in \mathrm{R}, i = 1, 2, 3, 4$ denote the angular speed of the rotors. For the design procedure, we give the notations as follow:

$$\boldsymbol{p} = [\,p_{x1} \quad p_{y1} \quad p_{z1} \quad p_{x2} \quad p_{y2} \quad p_{z2} \quad T_{\phi} \quad T_{\theta} \quad T_{\psi}\,]^{\mathrm{T}} \tag{3}$$

$$\boldsymbol{\Phi} \triangleq \boldsymbol{\Phi}(\dot{\boldsymbol{q}} \quad \ddot{\boldsymbol{q}}) = \begin{bmatrix} \ddot{\phi} & -\dot{\theta}\dot{\psi} & \dot{\theta}\dot{\psi} & & & & 1 & & \\ \dot{\phi}\dot{\psi} & \ddot{\theta} & -\dot{\phi}\dot{\psi} & & & & & 1 & \\ & & & -\dot{\phi}\dot{\theta} & \dot{\phi}\dot{\theta} & \ddot{\psi} & & & 1 \end{bmatrix} \tag{4}$$

Then Eq. (1) could be derived as:

$$\boldsymbol{U} = \boldsymbol{\Phi p} = \boldsymbol{\Phi}(\boldsymbol{p}_{\mathrm{n}} + \boldsymbol{p}_{\mathrm{u}}) \tag{5}$$

Where, $\boldsymbol{p}_{\mathrm{n}}$ is the parameter vector of nominal system, $\boldsymbol{p}_{\mathrm{u}}$ is the unknown parameter vector of the quadrotor UAV dynamic model.

2.2 New Version of Barbalate's Lemma

Theorem 1. If $x(t)$ is a uniformly continuous function and if there exist a lower bounded scalar function V and a continuous positive definite and radially unbounded scalar function M such that $-\dot{V} \geq M(x(t))$, then $x(t) \to 0$ as $t \to \infty$ [10].

3 Adatpive Nonsmooth Attitude Tracking Control (ANSATC) of Quadrotor UAV

Assumption 1. The reference attitude trajectory $\boldsymbol{q}_{\mathrm{r}} \in C^2$, and the sample rate of the controller is fast enough, so $\dot{\boldsymbol{p}} \approx 0$.

Backstepping approach will also be adopted to design the new adaptive nonsmooth attitude tracking controller (ANSATC) as follow.

Let $\boldsymbol{x}_1 \triangleq \boldsymbol{q}$ and $\boldsymbol{x}_2 \triangleq \dot{\boldsymbol{q}}$, lead to $\dot{\boldsymbol{x}}_1 = \boldsymbol{x}_2$. Define $\boldsymbol{e}_1 = \boldsymbol{x}_1 - \boldsymbol{x}_{1\mathrm{r}}$, $\boldsymbol{e}_2 = \boldsymbol{x}_2 - \boldsymbol{x}_{2\mathrm{r}}$ and $\boldsymbol{z} = \boldsymbol{x}_2 - \boldsymbol{\beta}$, where $\boldsymbol{x}_{1\mathrm{r}} = \boldsymbol{q}_{\mathrm{r}}$, $\boldsymbol{x}_{2\mathrm{r}} = \dot{\boldsymbol{q}}_{\mathrm{r}}$ are the desired trajectory, $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the tracking error of attitude and angular velocity respectively, and $\boldsymbol{\beta}$ is the auxiliary part which will be designed later.

Step 1. Consider the candidate Lyapunov function as follow

$$V_1 = \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{e}_1 / 2 \tag{6}$$

Differentiating V_1 versus time along the trajectories of system (1) and with the definitions given above yields

$$\dot{V}_1 = \boldsymbol{e}_1^{\mathrm{T}} \dot{\boldsymbol{e}}_1 = \boldsymbol{e}_1^{\mathrm{T}}(\boldsymbol{x}_2 - \boldsymbol{x}_{2\mathrm{r}}) = \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{z} + \boldsymbol{e}_1^{\mathrm{T}}(\boldsymbol{\beta} - \boldsymbol{x}_{2\mathrm{r}}) \tag{7}$$

Consider

$$\boldsymbol{z} = \boldsymbol{e}_2 - \boldsymbol{K}_1 \mathrm{sig}^{\alpha_1}(\boldsymbol{e}_1) \tag{8}$$

Where, $\boldsymbol{K}_1 \in \mathrm{R}_+^{3\times 3}$ is a diagonal positive-definite gain matrix, $\alpha_1 \in \mathrm{R}^+, 0 < \alpha_1 < 1$. And the notation $\mathrm{sig}^{\alpha_1}(z)$ is defined as

$$\mathrm{sig}^{\alpha_1}(z) \triangleq \left[\ |z_\varphi|^\alpha \mathrm{sign}(z_\varphi) \quad |z_\theta|^\alpha \mathrm{sign}(z_\theta) \quad |z_\psi|^\alpha \mathrm{sign}(z_\psi) \ \right]^{\mathrm{T}} \tag{9}$$

Where $\mathrm{sign}(\cdot)$ is a sign function [11]. From these equations, it is easy to deduce $\boldsymbol{\beta}$ as follow

$$\boldsymbol{\beta} = \boldsymbol{x}_{2\mathrm{r}} - \boldsymbol{K}_1 \mathrm{sig}^{\alpha_1}(\boldsymbol{e}_1) \tag{10}$$

Then Eq. (7) could be written as

$$\dot{V}_1 = \boldsymbol{e}_1^{\mathrm{T}} \boldsymbol{z} - \boldsymbol{K}_1 \boldsymbol{e}_1^{\mathrm{T}} \mathrm{sig}^{\alpha_1}(\boldsymbol{e}_1) \tag{11}$$

It is easy to deduce that $\dot{V}_1 = -\boldsymbol{K}_1 \boldsymbol{e}_1^{\mathrm{T}} \mathrm{sig}^{\alpha_1}(\boldsymbol{e}_1) \leq 0$ as $\boldsymbol{z} = 0$, which will be guaranteed by the designation below.

Step 2. Notating the system matrixes of the nominal system as $\boldsymbol{M}_{\mathrm{n}}$, $\boldsymbol{C}_{\mathrm{n}}$ and $\boldsymbol{T}_{\mathrm{n}}$ respectively, and consider that the nominal system is a desired system whose parameters are known and without external disturbances, which means $\boldsymbol{M}_{\mathrm{n}}$ is a diagonal positive-definite matric and $\boldsymbol{T}_{\mathrm{n}} = 0$. It is easily obtained that from (1).

$$\boldsymbol{U} = \boldsymbol{\Phi} \boldsymbol{p}_{\mathrm{u}} + \boldsymbol{\Phi} \boldsymbol{p}_{\mathrm{n}} = \boldsymbol{\Phi} \boldsymbol{p}_{\mathrm{u}} + \boldsymbol{M}_{\mathrm{n}} \ddot{\boldsymbol{q}} + \boldsymbol{C}_{\mathrm{n}}(\dot{\boldsymbol{q}}) \tag{12}$$

Then consider another Lyapunov function as follow

$$V = \left(\boldsymbol{z}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{n}} \boldsymbol{z} + \bar{\boldsymbol{p}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} \bar{\boldsymbol{p}}_{\mathrm{u}} \right)/2 \tag{13}$$

Where $\bar{\boldsymbol{p}}_{\mathrm{u}} \triangleq \boldsymbol{p}_{\mathrm{u}} - \widehat{\boldsymbol{p}}_{\mathrm{u}}$, $\widehat{\boldsymbol{p}}_{\mathrm{u}}$ is the estimation result of the adaption laws, which will be designed next. $\boldsymbol{\Gamma} \in \mathrm{R}_+^{3\times 3}$ is a diagonal positive-definite gain matrices.

Differentiating V with respect to time and yields

$$\dot{V} = \boldsymbol{z}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{n}} \dot{\boldsymbol{z}} + \bar{\boldsymbol{p}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} \dot{\bar{\boldsymbol{p}}}_{\mathrm{u}} \tag{14}$$

Substituting Eqs. (8), (10) and (12) into (14), we obtain

$$\dot{V} = \boldsymbol{z}^{\mathrm{T}} \boldsymbol{M}_{\mathrm{n}} \left(\dot{\boldsymbol{x}}_2 - \dot{\boldsymbol{\beta}} \right) + \bar{\boldsymbol{p}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} \dot{\bar{\boldsymbol{p}}}_{\mathrm{u}} = \boldsymbol{z}^{\mathrm{T}} \left[-\boldsymbol{C}_{\mathrm{n}}(\dot{\boldsymbol{q}}) - \boldsymbol{\Phi} \boldsymbol{p}_{\mathrm{u}} - \boldsymbol{M}_{\mathrm{n}} \dot{\boldsymbol{\beta}} + \boldsymbol{U} \right] + \bar{\boldsymbol{p}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} \dot{\bar{\boldsymbol{p}}}_{\mathrm{u}} \tag{15}$$

Design $\boldsymbol{U}$ as follow

$$\boldsymbol{U} = -\boldsymbol{K}_2 \cdot \mathrm{sig}^{\alpha_2}(\boldsymbol{z}) + \boldsymbol{M}_{\mathrm{n}} \dot{\boldsymbol{\beta}} + \boldsymbol{C}_{\mathrm{n}}(\dot{\boldsymbol{q}}) + \boldsymbol{\Phi} \hat{\boldsymbol{p}}_{\mathrm{u}} \tag{16}$$

Where $\boldsymbol{K}_2 \in \mathrm{R}_+^{3\times 3}$ is a diagonal positive-definite gain matrices, $\alpha_2 \in \mathrm{R}^+, 0 < \alpha_2 < 1$ is the order of fractional power function.

Then from (14) to (16), it can be derived as

$$\dot{V} = -\boldsymbol{K}_2 \cdot \boldsymbol{z}^{\mathrm{T}} \mathrm{sig}^{\alpha_2}(\boldsymbol{z}) + \left(\dot{\bar{\boldsymbol{p}}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} - \boldsymbol{z}_2^{\mathrm{T}} \boldsymbol{\Phi} \right) \bar{\boldsymbol{p}}_{\mathrm{u}} \tag{17}$$

Let $\dot{\bar{\boldsymbol{p}}}_{\mathrm{u}}^{\mathrm{T}} \boldsymbol{\Gamma} - \boldsymbol{z}_2^{\mathrm{T}} \boldsymbol{\Phi} = 0$, from Assumption 1 adaptation law can be derived [12]

$$\dot{\hat{\boldsymbol{p}}}_{\mathrm{u}}^{\mathrm{T}} = -\boldsymbol{z}_2^{\mathrm{T}} \boldsymbol{\Phi} \boldsymbol{\Gamma}^{-1} \tag{18}$$

Then we get $\dot{V} \leq 0$. For the system analysis, we define a scalar function $N(\boldsymbol{z})$ as

$$N(\boldsymbol{z}) = \boldsymbol{K}_2 \cdot \boldsymbol{z} \cdot \mathrm{sig}^{\alpha_2}(\boldsymbol{z}) \tag{19}$$

With the results from (13) to (17), it is easily to deduce that the infimum of V ($\inf\{V\}$) exists, and supremum of V satisfies $\sup\{V\} \leq V(0)$, so $\boldsymbol{z}$, $\bar{\boldsymbol{p}}_{\mathrm{u}}$ should be bounded. From (8), (12) and (16), we can obtain

$$\boldsymbol{M}_{\mathrm{n}} \dot{\boldsymbol{z}} = -\boldsymbol{K}_2 \cdot \mathrm{sig}(\boldsymbol{z}) + \boldsymbol{\Phi} \bar{\boldsymbol{p}}_{\mathrm{u}} \tag{20}$$

Then we can easily deduce that $\dot{\boldsymbol{z}}$ is bounded, so $\boldsymbol{z}$ should be uniformly continuous and $N(\boldsymbol{z})$ is radially unbounded. By theorem 1 we can obtain that $\boldsymbol{z} \to 0$ as $t \to \infty$, then from (8) we can obtain that $\boldsymbol{e}_1 \to 0$ and $\boldsymbol{e}_2 \to 0$.

Remark 1. When $\alpha_1 = 1$ and $\alpha_2 = 1$, the adaptive nonsmooth controller in (16) will become a type of normal adaptive controller, furthermore it will become a PD controller when $\alpha_1 = 1$, $\alpha_2 = 1$ and $\hat{\boldsymbol{p}}_{\mathrm{u}} = 0$.

Remark 2. Compared with the adaptive nonsmooth controller proposed in [8], it's easy to find that the new adaptive nonsmooth controller proposed in this paper is with simpler structure and much better flexibility for performance adjustment.

4 Simulation Results

The parameters of the quadrotor VAU are $m = 2\mathrm{kg}$, $l = 0.2\,\mathrm{m}$, $I_x = 1.25\,\mathrm{kg} \cdot \mathrm{m}^2$, $I_y = 1.25\,\mathrm{kg} \cdot \mathrm{m}^2$, $I_z = 2.5\,\mathrm{kg} \cdot \mathrm{m}^2$, $b = 2.98 \times 10^{-6}\,\mathrm{N} \cdot \mathrm{s}^2/\mathrm{rad}^2$, $d = 1.14 \times 10^{-7}\,\mathrm{N} \cdot \mathrm{s}^2/\mathrm{rad}^2$. And the sampling rate is 10 kHz. For simplicity, the altitude control is ignored, which means the quadrotor UAV is holding the altitude eternally. The reference attitude trajectory could be represented as $\varphi_{\mathrm{r}} = (\pi/6)\sin(2\pi)(\mathrm{rad})$, $\theta_{\mathrm{r}} = (\pi/6)\sin(2\pi)(\mathrm{rad})$, $\psi_{\mathrm{r}} = 0$.

And $\Delta_\phi = 10\% mgl/I_x$, $\Delta_\theta = 10\% mgl/I_y$, with the assumption that the parameters are unknown to controller and the nominal parameters are given as $k_{1\mathrm{n}} = 0.8k_1$, $k_{2\mathrm{n}} = 0.6k_2$, $I_{\mathrm{xn}} = 0.7I_{\mathrm{x}}$, $I_{\mathrm{yn}} = 0.7I_{\mathrm{y}}$, $I_{\mathrm{zn}} = 0.7I_{\mathrm{z}}$. Four experiments are designed with $\boldsymbol{K}_1 = 5 \times \mathrm{diag}(\,1 \quad 1 \quad 1\,)$, $\boldsymbol{K}_2 = 4 \times 10^6 \times \mathrm{diag}(\,1 \quad 1 \quad 1\,)$, and the different part of the experiments are shown as Table 1, which represent normal PD controller (Method 1), traditional nonsmooth controller (Method 2), normal adaptive controller (Method 3), and the proposed method (Method 4) respectively.

The simulation results are illustrated in Fig. 1. Where Fig. 1(a) and (b) presents the tracking results of ϕ, obviously the proposed method (Method 4) has the fastest transient and the best tracking performances. And the tracking results of θ (in Fig. 1(c) and (d)) and ψ (in Fig. 1(e) and (f)) show the same result. Furthermore Fig. 2 gives the control input of the four experiments, which are all continuous without chattering.

Table 1. Table of parameters of controller.

Exp.	Parameters of controller		
1	$\alpha_1 = 1$	$\alpha_2 = 1$	$\Gamma^{-1} = 0 \times \mathrm{diag}(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$
2	$\alpha_1 = 0.9$	$\alpha_2 = 0.9$	$\Gamma^{-1} = 0 \times \mathrm{diag}(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$
3	$\alpha_1 = 1$	$\alpha_2 = 1$	$\Gamma^{-1} = 5 \times 10^6 \times \mathrm{diag}(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$
4	$\alpha_1 = 0.8$	$\alpha_2 = 0.8$	$\Gamma^{-1} = 5 \times 10^6 \times \mathrm{diag}(1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1)$

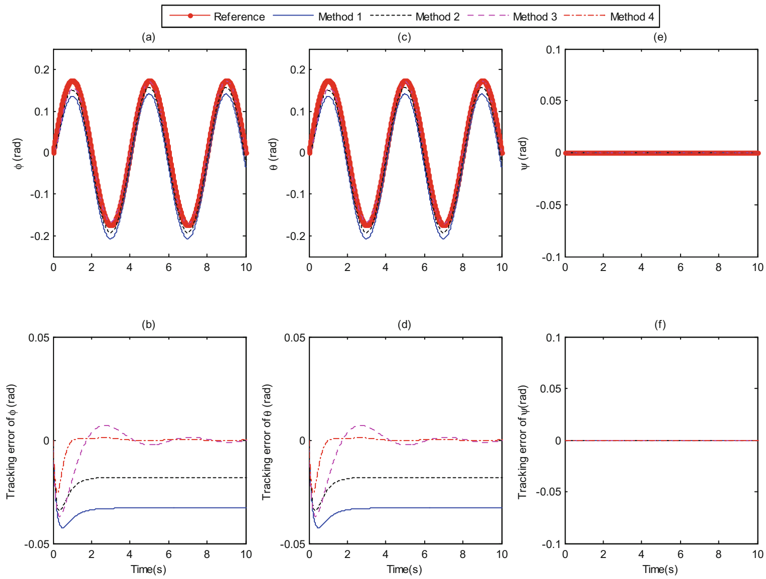


Fig. 1. Simulations results

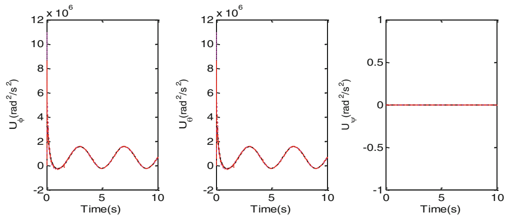


Fig. 2. Comparison of control input

Further, to evaluate the relationship between tracking performance and control parameter α_1 and α_2, four experiments are carried out with $\alpha_1 = 0.8$, $\alpha_2 = 0.8$, $\alpha_1 = 0.7$, $\alpha_2 = 0.7$, $\alpha_1 = 0.5$, $\alpha_2 = 0.7$ and $\alpha_1 = 0.5$, $\alpha_2 = 0.5$. Figure 3 illustrates the tracking results of ϕ, from Fig. 3(a) we can easily find that the tracking performance is best when choosing $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$, and from the enlarged drawings in Fig. 3(b), we can find high frequency chatting in U_ϕ when $\alpha_1 = 0.5$ and $\alpha_2 = 0.5$. To summarize, the attitude tracking performance can be improved by reducing α_1 and α_2, but α_1 and α_2 could not be arbitrary small. Because it may lead to chatting in real system, which may motivate the unmodeled system dynamics. When we choose $\alpha_1 = 0.5$, $\alpha_2 = 0.7$, the tracking performance is acceptable and without chatting problem, although it's not as good as the best one.

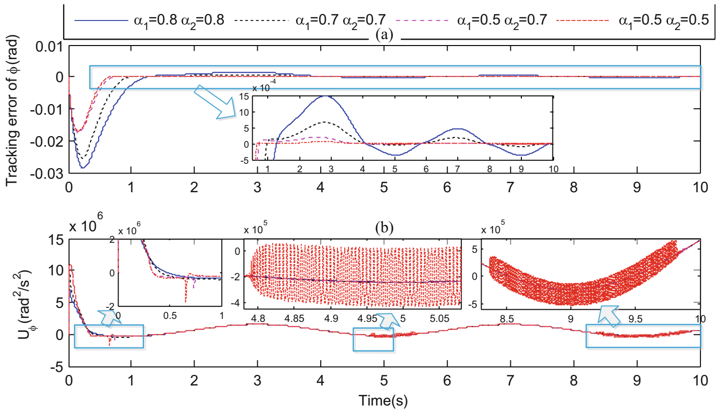


Fig. 3. Tracking result comparison of ϕ with different α_1 and α_2

5 Conclusion

In this paper, an adaptive nonsmooth attitude tracking control of quadrotor UAV with new structure is presented. It is designed in detail based on the dynamic model considering the dynamic uncertainties, and the system stability is analyzed. Finally the effectiveness of the new ANSATC was validated via serial simulations, which showed that the new controller can guarantee good transient and precise tracking performance while rejecting dynamic uncertainties, and has simpler structures and adjustment flexibility for better performance.

Acknowledgements. The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported in part by Initial Scientific Research Fund in FJUT under Grant GY-Z13101, in part by the Natural Science Foundation of Fujian Province under Grant 2018J01640, and in part by the Education and research projects for young teachers in Fujian Provincial Education Department under Grant JAT170369. and in part by the Technology Development Base of Fujian Industrial Integration Automation Industries.

References

1. Bouabdallah, S., Murrieri, P., Siegwart, R.: Design and control of an indoor micro quadrotor. In: IEEE International Conference on Robotics and Automation, pp. 4393–4398 (2004)
2. Wang, L., Su, J.: Robust disturbance rejection control for attitude tracking of an aircraft. IEEE Trans. Control Syst. Technol. **23**(6), 2361–2368 (2015)
3. Yang, H., Cheng, L., Xia, Y., et al.: Active disturbance rejection attitude control for a dual closed-loop quadrotor under gust wind. IEEE Trans. Control Syst. Technol. **26**(4), 1400–1405 (2017)
4. Guo, Y., Jiang, B., Zhang, Y.: A novel robust attitude control for quadrotor aircraft subject to actuator faults and wind gusts. IEEE/CAA J. Autom. Sinica **5**(1), 292–300 (2018)
5. Islam, S., Liu, P.X., Saddik, A.E.: Robust control of four-rotor unmanned aerial vehicle with disturbance uncertainty. IEEE Trans. Ind. Electron. **62**(3), 1563–1571 (2015)
6. Alexis, K., Nikolakopoulos, G., Tzes, A.: On trajectory tracking model predictive control of an unmanned quadrotor helicopter subject to aerodynamic disturbances. Asian J. Control **16**(1), 209–224 (2014)
7. Yang, Y., Chen, Y., Tang, C., et al.: Quadrotor helicopters trajectory tracking with stochastic model predictive control. In: 2017 Australian and New Zealand Control Conference (ANZCC), pp. 166–171. IEEE (2017)
8. He, D., Gao, P., Liu, L., et al.: Adaptive nonsmooth attitude tracking control of quadrotor UAV with dynamic uncertainties. In: International Conference on Smart Vehicular Technology, Transportation, Communication and Applications, pp. 135–142. Springer, Cham (2017)
9. Tao, Z.: Adaptive sliding control based on a new reaching law. Control Decis. **31**(7), 1335–1338 (2016)
10. Hou, M., Duan, G., Guo, M.: New versions of Barbalat's lemma with applications. J. Control Theory Appl. **8**(4), 545–547 (2010)
11. Su, Y., Zheng, C.: Global finite-time inverse tracking control of robot manipulators. Robot. Comput. Integr. Manuf. **27**(3), 550–557 (2011)
12. Slotine, J.E., Li, W., Cheng, D.: Applied Nonlinear Control. Machine Press, Beijing (2006)

Geo-spatial Ontology Matching Through Compact Evolutionary Algorithm

Xingsi Xue[1,2,3,4](✉) and Jianhua Liu[1,2,3]

[1] College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, Fujian, China
jack8375@gmail.com, jhliu@fjnu.edu.cn
[2] Intelligent Information Processing Research Center, Fujian University of Technology, Fuzhou 350118, Fujian, China
[3] Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, Fujian, China
[4] Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China

Abstract. Geo-spatial ontologies can provide a formal description of concepts, relationships, activities, features and rules in GIS domain. However, simply use them only allows to partially solve semantic conflicts, and does not completely solve heterogeneity issues that are caused by themselves. Geo-spatial ontology matching technique can find the correspondences between semantic identical entities, and solve the heterogeneous problem between two geo-spatial ontologies. Be inspired by the successful application of Evolutionary Algorithm (EA) in instance matching domain, in this paper, it is utilized to match the heterogeneous geo-spatial ontologies. To reduce the runtime and memory consumption required by EA, a compact version of it is presented, which does not work on the whole population but a probability representation on it. In addition, a geo-spatial similarity measure is presented to determine the identical geo-spatial entities, and an optimal model is constructed for geo-spatial ontology matching problem. The experimental results show that cEA-based geo-spatial ontology matching technique can efficiently determine the alignment.

Keywords: Geo-spatial ontology matching
Compact evolutionary algorithm · OAEI

1 Introduction

Geo-spatial ontology is a new research branch at the frontier of geographic information science and knowledge engineering [8], which is a special case of ontology that describes geo-spatial information domain knowledge, helps to understand the semantics and therefore to intelligently locate and integrate geo-spatial information for a wide variety of tasks. Geo-spatial ontologies can provide a formal description of concepts, relationships, activities, features and rules in GIS domain

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 11–18, 2019.
https://doi.org/10.1007/978-3-030-04585-2_2

[2], but simply use them only allows to partially solve semantic conflicts, and does not completely solve heterogeneity issues that are caused by themselves. To simplify and enhance the sharing, reuse and integration of geographic data and further implement the semantic inter-operability between geo-ontologies, the correspondences between semantic identical entities needs to be determined through geo-spatial ontology matching technique. A correspondence set between two geo-spatial ontologies is called ontology alignment, and each correspondence inside is a 3-tuples $(e_1, e_2, =, confidence)$, where e_1 and e_2 are respectively the entities of two geo-spatial ontology, the relation of the correspondence is the equivalence (=), and $confidence$ is the trust extent of the equivalence holds between e_1 and e_2.

Currently, there are only a few geo-spatial ontology matchers, such as Silk [3], LIMES [6] and GeOnt [4]. Be inspired by the successful application of Evolutionary Algorithm (EA) in ontology matching domain [12], in this paper, we propose to use it to match the heterogeneous geo-spatial ontologies. To reduce the runtime and memory consumption required by EA, a compact version of it is presented, which does not work on the whole population but a probabilistic representation on it. In addition, a geo-spatial similarity measure is presented to determine the identical geo-spatial entities, and an optimal model is constructed for geo-spatial ontology matching problem.

The rest of the paper is organized as follows: Sect. 2 describes the similarity measure on geo-spatial entities; Sect. 3 defines the geo-spatial ontology matching problem; Sect. 4 presents cEA in details; Sect. 6 shows the experimental results; finally, Sect. 7 draws the conclusions.

2 Similarity Measure

The similarity measure on geo-spatial entities is the foundation of Geo-spatial ontology techniques. In a real-world situation, geo-spatial entity can use different properties to describe the geographical information. For this reason, there is a need for the similarity measure that works with unknown properties, which provides techniques for comparing in general the statements belonging to two entities. The comparison should be carried out between the property-value pairs, and the search for comparable properties is necessary, which are the ones sharing the same or possess similar names. As an example, many geo-spatial entities contain geo-coordinates expressed using the standard geo:long and geo:lat properties, which can be easily detected and compared. In this study, we first standardize the special property values, e.g. replace with "2010-11-19" the time stamp value "November 19, 2010" or "11/19/10". Then, the similarity value of two geo-spatial entities e_1 and e_2 is calculated as follows:

$$sim_e(e_1, e_2) = \frac{\sum_{i=1}^{f} \max_{j=1\cdots g}(sim_p(<p_{1i}, v_{1i}>, <p_{2j}, p_{2j}>)) + \sum_{j=1}^{g} \max_{i=1\cdots f}(sim_p(<p_{1i}, v_{1i}>, <p_{2j}, p_{2j}>))}{f+g} \quad (1)$$

where p_{1i} and v_{1i} are e_1's ith property and its value, respectively, p_{2j} and v_{2j} are respectively e_2's jth property and its value. Once the property-value pair to be compared are selected, e.g. $<p_1, v_1>$ and $<p_2, v_2>$, the values are compared using string similarity metrics, and the property-value pairs to be compared are given a score which takes into account both the property similarity and the value similarity, as follows:

$$sim_p(<p_1, v_1>, <p_2, v_2>) = sim(p_1, p_2) \times sim(v_1, v_2) \tag{2}$$

The similarity value of two strings is calculated by SMOA distance [9], which is the most performing string-based similarity measure in ontology matching domain, and a linguistic measure, which calculate a synonymy-based distance through Wordnet [5]. Given two strings s_1 and s_2, the similarity $sim(s_1, s_2)$ is calculated according to the following formula:

$$sim(s_1, s_2) = \begin{cases} 1, & \text{if two words are synonymous} \\ SMOA(s_1, s_2), & \text{otherwise} \end{cases} \tag{3}$$

Particularly, SMOA distance between two strings s_1 and s_2 can be defined as $SMOA(s_1, s_2) = Comm(s_1, s_2) - Diff(s_1, s_2) + WinklerImpr(s_1, s_2)$, where $Comm(s_1, s_2)$ stands for the commonality between s_1 and s_2, $Diff(s_1, s_2)$ for the difference, and $WinklerImpr(s_1, s_2)$ for the improvement of the result from using the method introduced by Winkler [10].

3 Geo-spatial Ontology Matching Problem

Normally, the ontology alignment is assessed on the basis of three measures commonly known as: (1) recall, which determines whether all correct correspondences have been determined (completeness); (2) precision, which determines whether all determined correspondences are correct correspondences (soundness); and (3) f-measure, which is the harmonic mean of recall and precision [7]. However, it requires that the perfect matching result, i.e. the reference alignment, should be given in advance, which is generally unknown for practical applications. Therefore, supposing one entity in source ontology is matched with only one entity in target ontology and vice versa, in this paper, we use the rough alignment evaluation metrics, i.e., MatchCoverage, MatchRatio and MatchFmeasure [11], to respectively approximate the recall, precision and f-measure.

MatchCoverage is the fraction of those entities with at least one correspondence in the resulting alignment to the total number of entities in the ontology, which is used to approximate recall:

$$MatchCoverage = \frac{|C_{O_1-Match}| + |C_{O_2-Match}|}{|C_{O_2}| + |C_{O_2}|} \tag{4}$$

where:

- $C_{O_1-Match}$ and $C_{O_2-Match}$ are the set of matched entities of ontologies O_1 and O_2 respectively;
- C_{O_1} and C_{O_2} are the set of all entities of ontologies O_1 and O_2 respectively.

MatchRatio is the ratio between the number of found correspondences and the number of matched entities, which is used to approximate precision:

$$MatchRatio = \frac{|C_{O_1-Match}| + |C_{O_2-Match}|}{2 \cdot |Corr_{O_1-O_2}|} \tag{5}$$

where:

- $Corr_{O_1-O_2}$ is the set of correspondences in a resulting alignment from ontology O_1 to ontology O_2;
- $C_{O_1-Match}$ and $C_{O_2-Match}$ are the sets of matched entities of ontologies O_1 and O_2 respectively.

MatchFmeasure is the harmony mean of MatchCoverage and MatchRatio, which is used to approximate f-measure.

$$MatchFmeasure = 2 \times \frac{MatchCoverage \times MatchRatio}{MatchCoverage + MatchRatio} \tag{6}$$

On this basis, the geo-spatial ontology matching problem is defined as follows:

$$\begin{cases} max \quad f(X) \\ s.t. \quad X = (x_1, x_2, \cdots, x_{|O_1|})^T \\ \qquad\quad x_i \in \{0, 1, 2, \cdots, |O_2|\}, i = 1, 2, \cdots, |O_1| \end{cases} \tag{7}$$

where $|O_1|$ and $|O_2|$ respectively represent the cardinalities of source class set O_1 and target class set O_2, and $x_i, i = 1, 2, \cdots, |O_1|$ represents the ith pair of correspondence. In particular, $x_i = 0$ means the ith source class is mapped to none, and the objective function calculates the MatchFmeasure of a solution X's corresponding alignment.

4 Compact Evolutionary Algorithm

cEA simulates the behaviour of population-based EA by employing the probabilistic representation of the population, and a run of cEA is able to highly improve the performance of traditional EA in terms of both runtime and memory consumption. In the next, two main components of cEA, i.e. chromosome encoding mechanism and probability vector, are presented.

4.1 Chromosome Encoding Mechanism

In this work, the genes are encoded through the binary coding mechanism to represent the correspondences in the alignment. Given the total number of classes in source ontology and target ontology O_1 and O_2, the first part of a chromosome (or PV) consists of O_1 gene segments, and the Binary Code Length (BCL) of each gene segment is equal to $\lfloor \log_2(O_2) + 0.5 \rfloor$, which ensures each gene segment could present any target ontology class's index. While, the second part of a chromosome (or PV) has only one gene segment, whose BCL is equal to $\lfloor \log_2(\frac{1}{numAccuracy})+0.5\rfloor$, which can ensure this gene segment could present any threshold value under the numerical accuracy $numAccuracy$. Thus, the total length of the chromosome (or PV) is equal to $n_s \times \lfloor \log_2(n_t)+0.5\rfloor + \lfloor \log_2(\frac{1}{numAccuracy})+0.5\rfloor$. Given a gene $gene = \{geneBit_1, geneBit_2, \cdots, geneBit_n\}$ where $geneBit_i$ is the ith gene bit value of the gene segment, we decode it to obtain a decimal number whose value is equal to $\sum_{i=1}^{n} 2^{geneBit_i}$. In particular, the decimal numbers obtained represent the indexes of the target classes, where 0 means the source instance is not mapped to any target ontology's class.

4.2 Probability Vector

In this work, we use one Probability Vector (PV) to characterize the entire population. The number of elements in PV is equal to the number of individual's gene bits and each element's value is in [0,1]. Since each element's value in PV represents the probability of being one, we can use PV to generate various solutions. In addition, PV can be updated based on the better solution in terms of its fitness value, with the aim to move the PV toward the better solution. Here is an example of generating a new solution through PV $(0.1, 0.8, 0.5, 0.9)^T$. First, generate four random numbers, such as 0.4, 0.5, 0.8 and 0.1. Then compare the numbers with the elements in PV accordingly to determine the new generated individual's gene values, For example, since $0.4 > 0.1$, the first gene bit's value of the new solution is 0, and similarly, the remaining gene bits' values are 1, 0 and 1, respectively. In this way, the new solution we obtain is 0101. By repeating this procedure, we can obtain various individuals. In addition, if 0101 is the best solution in the current generation, i.e. elite, PV should be updated according to its information. Given the PV update value $updateValue$, say 0.1, if the gene value of the elite is 0, the corresponding element of PV will decrease by $updateValue$, otherwise increase by $updateValue$. In this way, the updated PV is $(0.0, 0.9, 0.4, 1.0)^T$.

5 Pseudo-code of Compact Evolutionary Algorithm

The pseudo-code of cEA is given as follows:

Step (1) Initialization:

1. generation=0;
2. for($i = 0$; $i < PV.length$; i++)
3. $PV[i] = 0.5$;
4. end for
5. generate an individual ind_{elite} through PV;

Step (2) Update PV:

```
generate an individual ind_new through PV;
[winner, loser] = compete(ind_elite, ind_new);
if (winner == ind_new)
  ind_elite = ind_new;
end if
for(i = 0; i < solution.length; i++)
  if (winner[i]==1)
    PV[i] = PV[i] + 0.1;
    if (PV[i] > 1)
      PV[i] = 1;
    end if
  else
    PV[i] = PV[i] − 0.1;
    if (PV[i] < 0)
      PV[i] = 0;
    end if
  end if
end for
if (each bit of PV is either 1 or 0)
  for(i = 0; i < PV.length; i++)
    PV[i] = 0.5;
  end for
end if
```

Step (3) Stopping Criteria:

```
if (maxGeneration = 3000 is reached)
  stop and output ind_elite;
else
  generation=generation+1;
  go to Step 2;
end if
```

6 Experimental Results

In the experiment, the Linking benchmark in HOBBIT Link Discovery track provided by the well-known Ontology Alignment Evaluation Initiative (OAEI 2017)[1] is utilized to test our proposal's performance, which measures how well the matchers can match traces that have been modified using string-based approaches along with addition and deletion of intermediate points. Linking benchmark focuses on string-based transformations with different levels, types of spatial object representations and types of date representations. OAEI's participants' results are from [1], and our results are the average values over 30 independent runs.

[1] https://project-hobbit.eu/challenges/om2017/.

Table 1. Comparison with OAEI 2017's participants on Linking benchmark.

Matcher	Recall	Precision	f-measure	Runtime (second)
AML	1.00	1.00	1.00	11,722
OntoIdea	0.99	.99	.99	19,806
cEA	1.00	1.00	1.00	9,762

Finally, as can be seen from Table 1, cEA's result is equal to or better than the participants of OAEI 2017 in terms of f-measure, and the runtime is the lowest, which shows the effectiveness of our approach.

7 Conclusion

Geo-spatial ontology matching can find the correspondences between semantically identical geo-spatial entities, and the alignment determined can be further utilized to implement the semantic inter-operability between geo-ontologies. In this paper, a cEA-based geo-spatial ontology matching technique is proposed to automatically determine the high quality geo-spatial ontology alignment. The experimental results show that our proposal can efficiently determine the alignment.

Acknowledgment. This work is supported by the National Natural Science Foundation of China (No. 61503082), Natural Science Foundation of Fujian Province (No. 2016J05145), Scientific Research Development Foundation of Fujian University of Technology (Nos. GY-Z17162 and GY-Z15007) and Fujian Province Outstanding Young Scientific Researcher Training Project (No. GY-Z160149).

References

1. Achichi, M., et al.: Results of the ontology alignment evaluation initiative 2017. In: OM 2017-12th ISWC Workshop on Ontology Matching, pp. 61–113 (2017). No commercial editor
2. An, Y., Zhao, B.: Geo ontology design and comparison in geographic information integration. In: 2007 Fourth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2007, vol. 4, pp. 608–612. IEEE (2007)
3. Isele, R., Jentzsch, A., Bizer, C.: Efficient multidimensional blocking for link discovery without losing recall. In: WebDB (2011)
4. Jaziri, W., Chaabane, S., Sassi, N.: GeOnt: geo-ontologies integration tool. Int. J. Metadata Semant. Ontol. **12**(2–3), 155–166 (2017)
5. Miller, G.A.: WordNet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
6. Ngomo, A.C.N.: Orchid-reduction-ratio-optimal computation of geo-spatial distances for link discovery, pp. 395–410. Springer (2013)
7. Rijsberge, C.J.V.: Information Retrieval. University of Glasgow, Butterworth (1975)

8. Siricharoen, W.V., Pakdeetrakulwong, U.: A survey on ontology-driven geographic information systems. In: 2014 Fourth International Conference on Digital Information and Communication Technology and it's Applications (DICTAP), pp. 180–185. IEEE (2014)
9. Stoilos, G., Stamou, G., Kollias, S.: A string metric for ontology alignment, pp. 623–637, Galway, Ireland, November 2005
10. Winkler, W.: The state record linkage and current research problems. Technical report. RR99-04, Statistics of Income Division, Washington DC, USA (1999)
11. Xue, X., Wang, Y.: Optimizing ontology alignments through a memetic algorithm using both matchfmeasure and unanimous improvement ratio. Artif. Intell. **223**, 65–81 (2015)
12. Xue, X., Wang, Y.: Using memetic algorithm for instance coreference resolution. IEEE Trans. Knowl. Data Eng. **28**(2), 580–591 (2016)

Cryptanalysis and Improvement of a Remote Three-Factor Authentication Protocol for the Multi-server Environment

Chien-Ming Chen[1], Yanyu Huang[1], Xiaoting Deng[1], and Tsu-Yang Wu[2,3,4](✉)

[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China
chienming.taiwan@gmail.com, 736096421@qq.com, 1553968578@qq.com

[2] College of Computer Science and Engineering, Shandong University of Technology, Shandong, China
wutsuyang@gmail.com

[3] Fujian Provincial Key Laboratory of Big Data Mining and Applications, Fuzhou, China

[4] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education (Fujian University of Technology), Fujian University of Technology, Fuzhou 350118, China

Abstract. Nowadays, a multi-server environment has been widely used in various applications. Under this environment, users can register in one particular register center just one time and log into different application servers. Recently, various authentication protocol for this kind of environment have been proposed. In 2017, Zhang et al. proposed a three-factor authentication protocol with strong robustness. They claimed that their protocol can secure against various kinds of attacks, including perfect forward secrecy. Unfortunately, in this paper, we demonstrated that Zhang et al.'s protocol still cannot provide Perfect Forward Secrecy. To erase the weakness, we proposed an improvement to let the protocol can provide Perfect Forward Secrecy.

1 Introduction

The rapid development of internet technologies and mobile devices has greatly facilitated our lives, but at the same time, the traditional single-server environment exposed some weaknesses or inconveniences. For example, users should register with different application server and remember different identity and passwords, which is inconvenience and impractical to users. As a result, multi-server environment has been designed to solve the problems. In multi-server environment, users can register in a public register center only once and log into different application servers with same identity and passwords.

In order to provide mutual authentication and generate a session key, lots of authenticated and key agreement (AKA) protocols [1–3,11–13] for different kinds of applications and environments have been proposed. There are also lots

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 19–24, 2019.
https://doi.org/10.1007/978-3-030-04585-2_3

of AKA protocols [5–8,10] are designed for multi-server environment. In 2014, He et al. [4] proposed a three-factor authentication protocol based on Elliptic Curve Cryptographic (ECC) and smart card. However, Odelu et al. [9] then demonstrated that He et al.'s protocol is vulnerable to a known session-specific temporary information attack and an impersonation attack. Odelu et al. came up with a modification and claimed that they fixed the weaknesses they found. However, in 2017, Zhang et al. [14] pointed out that Odelu et al.'s protocol is still suffers from a denial of service (DoS) attack and an insider attack. Besides, it cannot provide strong robustness. Then Zhang et al. proposed another authentication protocol. They claimed that their protocol can secure against various kinds of attacks, including the perfect forward secrecy (PFS). If an AKA protocol can provide the perfect forward secrecy, an attacker cannot calculate the session key even if this attacker obtains all secrets of all participants. However, in this paper, we demonstrate that Zhang et al.'s protocol still does not provide Perfect Forward Secrecy. To erase the weakness, we proposed our improvement to provide the perfect forward secrecy.

The reminder of the paper is organized as follows. Section 2 reviews and analyses Zhang et al.'s Protocol. In Sect. 3, we demonstrate that Zheng et al.'s protocol does not provide perfect forward secrecy. Then we will propose our improvement in Sect. 4. finally, Sect. 5 concludes.

2 Review of Zhang et al.'s Protocol

In this section, we briefly review the Zhang et al.'s protocol. This protocol contains the following four phases, the initialization phase, the registration phase, the login and authentication phase, and the password change phase. In this paper, we only describe the former three phases. The detailed steps of Zhang et al.'s protocol can refer to their paper.

Initialization Phase: In this phase, Register Center (RC) first selects a private key k and a random seed x and then generates Chebyshev Chaotic Map with k and x. After that, RC publishes $\{T_k(x), x\}$ and keeps k securely.

Registration Phase: Both an application server and users need to register with RC. If a new application server desires to register with RC, it will perform the following steps. Note that all communication in this phase are through a secure channel.

1. If a new application server S_j wants to register with RC, S_j selects an identity SID_j and current timestamp t_i. It then sends SID_j and t_i to RC.
2. After receiving SID_j and t_i, RC first checks the validity of t_i by checking whether t_i exceeds the maximum range $\triangle t$. If this holds, RC calculates the private key s and public key PK_{S_j} of S_j with RC's private key k and current timestamp t_j. Note that $s = h(SID_j||k||t_j)$ and $PK_{S_j} = T_s(x)$. RC then transmits $\{s, PK_{S_j}\}$ back to S_j.

3. Now S_j keeps s as a secret and publishes $\{SID_j, PK_{S_j}, t_j\}$.

Besides, if an user wants to register with RC, he will perform the following steps. Also note that all communication is this phase are through a secure channel.

1. User U_i selects parameters $\{ID_i, pw_i, B_i^{reg}, t_{reg}^u\}$ to RC, which respectively represent identity, password, biometric features and current timestamp.
2. U_i calculates R_{ss} $=SS(B_i^{reg})$ and RPW_i $=h(ID_i||pw_i||B_i^{reg})$ and sends $\{ID_i, RPW_i, h(pw_i||B_i^{reg}), t_{reg}^u\}$ to RC.
3. After receiving the above messages from U_i, RC checks the validity of the t_{reg}^u and if it holds, RC uses k to calculates the private key u and public key PK_{U_i} of U_i. Note that $u = SK_{U_i} = h(ID_i||k||t_{reg}^u)$ and $PK_{U_i} = T_u(x)$. RC also computes $Z_i = u \oplus RPW_i$, X_i $=h(u||h(PW_i||B_i)||ID_i)$. Then RC inserts $\{X_i, Z_i, h(), p, x\}$ into smart card and sends the smart card back to U_i.
4. User add the R_{ss} into smart card.

2.1 Login and Authentication Phase

1. U_i inserts his smart card into the terminal and enters his identity ID_i, password pw_i and imprints U_i's biometric B_i^{reg}. Besides, U_i selects SID_j, public key PK_{S_j}, registration time t_j of the application server which U_i wants to visit.
2. Then, U_i recovers B_i^{reg} by SS algorithm and computes $RPW_i = h(ID_i|| pw_i||B_i^{reg})$, $u = RPW_i \oplus Z_i$, $PK_{U_i} = T_u(x)$, $X_i' = h(u||h(pw_i||B_i^{reg})||ID_i)$. U_i then compares if $X_i' = X$. If this holds, U_i generates random nonce a, computes $T_a(x)$, $T_uT_s(x)$, $T_aT_s(x)$, $M_1 = ID_i \oplus h(SID_j||T_aT_s(x))$, $M_2 = PK_{U_i} \oplus h(T_aT_s(x)||t_{login}$
$||SID_j)$ and $M_3 = h(ID_i||SID_j||T_aT_s(x)||T_uT_s(x)||T_a(x))$. Next, U_i transmits $\{T_a(x),\ M_1,$
$M_2,\ M_3,\ t_{login}\}$ to server S_j.
3. After the messages $\{T_a(x),\ M_1,\ M_2,\ M_3,\ t_{login}\}$ arrive, S_j first checks the validity of t_{login} and if this holds, S_j calculates $T_aT_s(x)$, $ID_i = M_1 \oplus h(SID_j||T_sT_a(x))$, $PK_{U_i} = M_2 \oplus h(T_sT_a(x)||t_{login}||SID_j)$ $=T_u(x)$, $T_sT_u(x)$, and $M_4 = h(ID_i||SID_j||T_sT_a(x)||T_sT_u(x)||T_a(x))$. S_j then compares that whether M_4 is euqual to M_3. If this holds, S_j generates b and computes $T_b(x), T_bT_a(x)$, the session key $SK_{ji} = h(ID_I||SID_j||T_sT_a(x)|| T_sT_u(x)||T_bT_a(x))$ and $M_5 = h(ID_i||SK_{ji}||T_sT_u(x)||T_bT_a(x)||T_b(x))$. At last, S_j transmits $\{T_b(x), M_5\}$ back to the U_i.
4. If U_i receiving the messages from S_j, U_i computes $T_aT_b(x)$, $SK_{ij}' = h(ID_i|| SID_j||\ T_aT_s(x)||T_uT_s(x)||T_aT_b(x))$, and $M_6 = h(ID_i||SK_{ij}'||T_uT_s(x)||T_a T_b(x)||T_b(x))$. Finally, U_i verifies that if M_5 is equal to $M6$. If this holds, U_i then computes $M_7 = h(T_uT_s(x)||T_aT_b(x)||T_aT_s(x))$ and sends M_7 to S_j.
5. S_j computes $M_8 = h(T_sT_u(x)||T_bT_a(x)||T_sT_a(x))$ and compares that if M_7 is equal to M_8. If this holds, the authentication progress is complete and successful, S_j and U_i can securely communicate using session key SK_{ij} and SK_{ji}.

3 Cryptanalysis of Zhang et al.'s Scheme

In this section, we demonstrate that Zhang et al.'s scheme does not provide the perfect forward secrecy.

3.1 The Adversary Model

We assume that the adversary E has the following capacities.

1. E has limited/compeleted control over the messages that were transmitted over the insecure channel, such as intercepting, modifying and deleting the transmitted message.
2. E can extract the security parameters stored in the smart card by using some power analysis technique.
3. E can try to obtain sensitive information(e.g. password) by using off-line password guessing attacks.

3.2 Perfect Forward Secrecy

Here we show that Zhang et al.'s scheme does not provide the perfect forward secrecy. More specifically, if an attacker E obtains transmitted messages over the insecure channel, the private key s of S_j and the security information in the smart card, E can compute the session key.

The final mission of E is to obtain the session key. As mentioned above, $SK_{ij} = SK_{ji} = h(ID_i||SID_j||T_sT_a(x)||T_sT_u(x)||T_bT_a(x))$. The detailed steps are listed as follows.

1. ID_i is equal to $M_1 \oplus h(SID_j||T_sT_a(x))$. Since SID_j is a public message and $M_1, T_a(x)$ is transmitted over a normal channel, E can computes $T_sT_a(x)$ with secret key s of S_j. Of course, E can get ID_i now.
2. Because E knows the M_2, $T_aT_s(x)$, t_{login}, $SID_j(T_aT_s(x) = T_sT_a(x))$, so E can get $T_u(x) = PK_{U_i}$ by $M_2 = PK_{U_i} \oplus h(T_aT_s(x)||t_{login}||SID_j)$.
3. E can obtain $T_aT_b(x)$ by $M_7 = h(T_uT_s(x)||T_aT_b(x)||T_aT_s(x))$.
4. As E can obtain the sensitive parameters above, the attacker can obtain session key by $SK_{ij} = SK_{ji} = h(ID_i||SID_j||T_sT_a(x)||T_sT_u(x)||T_bT_a(x))$, which means Zhang et al.'s scheme does not provide perfect forward secrecy.

4 Possible Improvement

According to our observation, the reason that this scheme does not provide the perfect forward secrecy is this scheme add the $T_aT_b(x)$ into the transmitted parameter M_5 and M_7. It lets an attacker can guessing the $T_aT_b(x)$. In order to solve the problem, we improve the calculation of M_5 and M_7. More specifically, M_5 is changed to $h(ID_i||T_sT_u(x)||T_b(x))$ and M_7 is changed to $h(T_uT_s(x)||T_aT_s(x)||T_a(x))$.

Now we can find that an attacker cannot obtain the $T_aT_b(x)$ through our attack mentioned above. Without the value of $T_aT_b(x)$, an attacker cannot further compute the session key. In conclude, our improvement can provide the perfect forward secrecy.

Furthermore, we also compare our improvement with related protocols in Table 1.

Table 1. Comparisons of the security features among protocols

SF	Protocols		
	Odelu and Das et al.'s	Zhang et al.'s	Our modified
A1	Yes	Yes	Yes
A2	Yes	Yes	Yes
A3	No	Yes	Yes
A4	No	Yes	Yes
A5	No	No	Yes

Abbreviations: **SF**: security features; **A1**: resist Password guessing attack; **A2**: resist Personation attack; **A3**: resist Insider attack; **A4**: resist DoS attack; **A5**: provide Perfect Forward Secrecy;

5 Conclusion

In this paper, we analyzed Zang et al.'s protocol. Although Zhang et al. claimed that their protocol can resist various kinds of attacks, we still find that their protocol cannot provide perfect forward secrecy. In order to erase the weakness, we proposed an improvement to guarantee the protocol is secure.

Acknowledgement. The work of Chien-Ming Chenwas supported in part by Shenzhen Technical Project under Grant number JCYJ20170307151750788 and in part by Shenzhen Technical Project under Grant number QJSCX20170327161755. The work of Tsu-Yang Wu was supported in part by the Science and Technology Development Center, Ministry of Education, China under Grant no. 2017A13025 and the Natural Science Foundation of Fujian Province under Grant no. 2018J01636.

References

1. Chen, C.M., Ku, W.C.: Stolen-verifier attack on two new strong-password authentication protocols. IEICE Trans. Commun. **85**(11), 2519–2521 (2002)
2. Chen, C.M., Xu, L., Wu, T.Y., Li, C.R.: On the security of a chaotic maps-based three-party authenticated key agreement protocol. J. Netw. Intell. **1**(2), 61–66 (2016)
3. He, B.Z., Chen, C.M., Wu, T.Y., Sun, H.M.: An efficient solution for hierarchical access control problem in cloud environment. Math. Probl. Eng. (2014)

4. He, D., Wang, D.: Robust biometrics-based authentication scheme for multiserver environment. IEEE Syst. J. **9**(3), 816–823 (2015)
5. Juang, W.S.: Efficient multi-server password authenticated key agreement using smart cards. IEEE Trans. Consum. Electron. **50**(1), 251–255 (2004)
6. Lee, C.C., Lin, T.H., Chang, R.X.: A secure dynamic ID based remote user authentication scheme for multi-server environment using smart cards. Expert Syst. Appl. **38**(11), 13863–13870 (2011)
7. Liao, Y.P., Wang, S.S.: Improvement of the secure dynamic ID based remote user authentication scheme for multi-server environment. Comput. Stand. Interfaces **31**(6), 1118–1123 (2009)
8. Liao, Y.P., Wang, S.S.: A secure dynamic ID based remote user authentication scheme for multi-server environment. Comput. Stand. Interfaces **31**(1), 24–29 (2009)
9. Odelu, V., Das, A.K., Goswami, A.: A secure biometrics-based multi-server authentication protocol using smart cards. IEEE Trans. Inf. Forensics Secur. **10**(9), 1953–1966 (2015)
10. Pan, J.S., Tso, R., Wu, M.E., Chen, C.M.: Security analysis of an anonymous authentication scheme based on smart cards and biometrics for multi-server environments, pp. 59–69. Springer (2015)
11. Sun, H.M., He, B.Z., Chen, C.M., Wu, T.Y., Lin, C.H., Wang, H.: A provable authenticated group key agreement protocol for mobile environment. Inf. Sci. **321**, 224–237 (2015)
12. Wang, K.H., Chen, C.M., Fang, W., Wu, T.Y.: A secure authentication scheme for internet of things. Perv. Mob. Comput. **42**, 15–26 (2017)
13. Wang, K.H., Chen, C.M., Fang, W., Wu, T.Y.: On the security of a new ultra-lightweight authentication protocol in IoT environment for RFID tags. J. Supercomput. **74**(1), 65–70 (2018)
14. Zhang, M., Zhang, J., Tan, W.: Remote three-factor authentication protocol with strong robustness for multi-server environment. China Commun. **14**(6), 126–136 (2017)

Research on Optimization of LMS Frequency Estimation Based on FPGA Technology

Liwen Chen(✉), Rijing Zheng, Yao Zhou, Wenji Zhang, and Renwu Yan

School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, Fujian, China
chenlw2002@whu.edu.cn

Abstract. The LMS algorithm (Least mean square) is common in the adaptive algorithm system. It has a simple calculation and convenient implementation. The convergence performance is excellent and the stability is high in the case of limited precision. The advantages of anti-interference on filtering are obvious. In this paper, the concept of LMS algorithm, the working principle and the selection of technical research scheme on frequency estimation are introduced, and we realize the LMS algorithm by FPGA technology and used it in clutter rejection area.

Keywords: LMS algorithm · Adaptive filter · FPGA · Frequency estimation

No matter what kinds of signals are transmitted, there are deviations and detuning because the process of transmission is in the transmission channels with external noise interferences. Therefore, the detection and treatment of signals are particularly important, it can issue the desired signal to be extracted from the complex signal in the condition of receiving, and inhibit the noise and interferences.

1 Introduction

An important content in the process of detection and treatment is to extract and restore the original signals from the jammers. To realize the function, one method is to design a filter. In the early 1940s, Wiener established the Wiener filter theory and the Wiener Filter was designed. When the signal is transmitted through the Wiener filter, the interference signal can be eliminated to the maximum extent and the desired signals can be reproduced and estimated as accurate as possible, and the noise associated with the signal is suppressed, too [1]. However, its effect will be greatly reduced as the input signal is limited and its statistical characteristics deviate from its original design conditions. In the 1960's, Kalman found a recursive filtering algorithm that can be realized by digital computer, which was applied in the fields of aviation and industry areas successfully. Later, it is called Kalman filter algorithm [2]. With the development of science and technology, and complexity and diversification of applications, and the uncertainty increasing of signals, and its own filtering restrictions are more harsh, it needs to know the system state equations and measurement equations in the actual

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 25–32, 2019.
https://doi.org/10.1007/978-3-030-04585-2_4

process, at the same time, the signals are all in the input status. However, it is difficult to obtain the state of the desired precisely, and it is also hard to be counted [3]. The demand of filter technique is getting higher for the development of technology and the signal processing technique is used in various fields. Based on the previous algorithm and the actual demand, the adaptive filter is generated through its own iteration to complete the adjustment of their own parameters to make certain theories come true, so as to achieve the best filter [4]. The adaptive filter is studied in different application fields, and achieves a lot of new results. For example, Howells has invented the intermediate frequency sidelobe cancellation by adaptive antenna technique. Widrow and Hoff optimized the LMS (least mean square) algorithm when studying the pattern recognition scheme of adaptive elements based on above basic research results. They provide the theoretical basis for the later adaptive filtering algorithm [5]. Since then, a wide variety of adaptive filtering algorithms are proposed, and constantly improve the capability of the adaptive LMS algorithms. Such as data block LMS algorithm, RLMS algorithm and so on. Because LMS algorithm has the advantages of low computational complexity, easy to use, good stability, and the combination of many filtering algorithms, it is better to be applied in various fields [6]. The mainly applies of LMS algorithm as follows: Echo cancellation [7], adaptive equalization of channels, Speeches from Social Media [8], system identification, wireless sensor network [9], adaptive beamforming, broadband Light [10], linear prediction and so on.

In this paper, we will research the frequency estimation technology of LMS algorithm and its FPGA implementation, which can be applied in the actual projects. In recently years, with the rapid development of digital signal technology and integrated circuit technology, many intelligent chips appear with excellent performance, and the improvement of general computer in related fields makes the adaptive LMS algorithm a solid hardware foundation. Therefore, the LMS algorithm will have a wider application prospect in the future.

2 LMS Algorithm Principle

2.1 Introduction of Adaptive LMS Algorithm

We cannot determine the statistical characteristics of the input signals change in the actual processes and the adaptive filtering algorithm can adjust its parameters in time to achieve the standards. Figure 1 is its schematic diagram.

The diagram below shows the diagram of the LMS algorithm. We can see that the LMS algorithm includes two main processes: filter processing and adaptive adjustment. At the n time, $x(n)$ is the input signal, $y(n)$ is the output signal, $d(n)$ is the expected signal. The LMS adaptive filter parameters are automatically adjusted the difference between the desired signal and the output signal, and the output $y(n + 1)$ of the next moment can close to the desired signal.

By comparing and analyzing the expected signal and the output signal, we can estimate the error as follows:

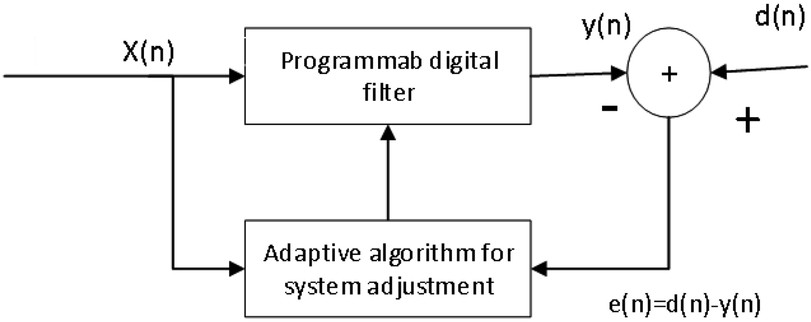


Fig. 1. Adaptive filter principle diagram

$$e(n) = d(n) - y(n) = d(n) - W^T(n)X(n) \tag{1}$$

The mean square error is:

$$J = E[e^2(n)] = E[d^2(n)] - 2E[d(n)W^T(n)X(n)] + E[W^T(n)X(n)X^T(n)W(n)] \tag{2}$$

Using the steepest descent algorithm and hunting the minimum point, the filter weight vector is adjusted along the steepest descent direction (negative gradient direction) on the performance surface. The iterative formula of calculating the weight vector is as follows:

$$W(n+1) = W(n) + \mu(-\nabla J) \tag{3}$$

μ is Step Length Factor.

In the steepest descent algorithm, we need to obtain the relevant information of the input signal and the output signal, and expect to get the best Wiener solution. In our calculation, the expected signal is unknown and, the correlation is unidentified. We have to estimate the gradient vector. The adaptive LMS algorithm uses the transient mean square error directly to calculate the gradient of the instantaneous tap vector:

$$\hat{\nabla} J = \frac{\partial(e^2(n))}{\partial W(n)} = [2e(n)\frac{\partial e(n)}{\partial w_1(n)}, 2e(n)\frac{\partial e(n)}{\partial w_2(n)}, \cdots, 2e(n)\frac{\partial e(n)}{\partial w_M(n)}]^T = -2e(n)X(n) \tag{4}$$

According to the above, the weight vector updating equation of adaptive LMS filtering algorithm is as follows:

$$W(n+1) = W(n) + \mu(-\nabla J) = W(n) + 2\mu\mu e(n)X(n) \tag{5}$$

3 Design Scheme and Demonstration

3.1 FPGA Design

FPGA (Field-Programmable Gate Array) is a field programmable gate array. There are a large number of logic gates and flip-flops in the FPGA chip with large scale and high integration, higher processing speed. It is a very practical design. FPGA has a good effect on some monotonous and complex algorithm implementation. In this paper, the FPGA hardware technology can easily complete parallel processing. You needn't use multiple IC chips to design a complex combinational circuit. A programmable design on a FPGA chip will do the trick, greatly shorten the cycle of project development.

3.1.1 Design of LMS Filter

In the design of the filter, the system parameters should be analyzed and determined. The analysis results of two important parameters in the LMS algorithm are given as follows.

One is the order of LMS filter. As the order of the equilibrium we choose is small, the performance is not very good, because the performance of the LMS algorithm is affected by the order. Conversely, we find the higher the order, the better the effect. But it takes up a lot of hardware resources. Therefore, the filter section with 8-order is chosen.

The second is step size factor. In the LMS algorithm, the greater the step factor, the faster the system converges at the cost of the larger steady-state error. However, using the smaller step factor, the running time of the system converges with the smaller the monostable error will become slower.

According to the continuous simulation verification, the step factor is $\mu = 0.001$.

Through the equalizer, the signal in the operation step passes through three parts, including filtering part, error analysis part, and filter coefficient updating part. Then the filtered signal is sent to the judgment module to recover the corresponding original signal.

The algorithm table for equalization filters is shown in Table 1.

Table 1. Flow chart of equalization filter algorithm

Step One	Initialize $\omega(0) = 0; L = 16; \mu = 0.001; i = 17$
Step Two	Update $i = i + 1;$ $for\ i = n + 1 : m$ $rr = r(i : -1 : i - n + 1)';$ $e(i) = (f'*rr) * (1 - (f'*rr)^\wedge 2);$ $f = f + mu * e(i) * rr;$ end
Step Three	Decision Output $dec = sign(y)$

3.1.2 Design of MIF Files

Throughing the MATLAB software program design, we can directly use MATLAB software to simulate the generated filter input signal before the next FPGA simulation. Firstly, using the data generated from the input signal and saving it in the mif file, and then we use it running in the Quartus II software. The generated mif file will be stored in ROM and then read it as the input signal data of the equalizer by using table technology.

3.2 FPGA Design of LMS Equalizer

3.2.1 Design of the Main Process

At the beginning, the filter response speed of LMS equalizer should be both fast and inexpensive. This design adopts 8-order filter. This can improve the speed of the operation greatly. The design is divided into 4 parts, plus an input storage module. The top level design of the LMS adaptive equalizer is shown in Fig. 2.

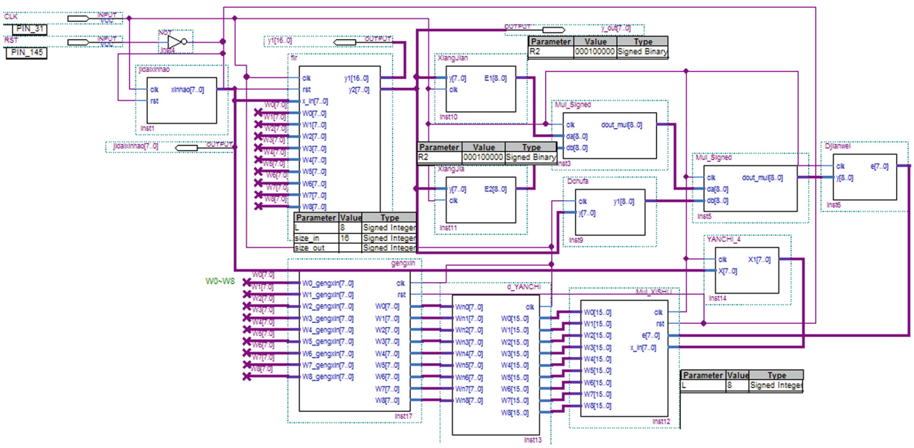


Fig. 2. Top level design of LMS adaptive equalization filter

3.2.2 Design of Filter Module

The function on the transversal filter module is to partially compensate the input signal. When we design the 8-bit signed number, the output signal can get a 17-bit signed number after convolution with its coefficient, we process it and turn it into 8 bits again and then sent it to the judgment module to determine the corresponding output signal. We use an 8-order transversal filter, *clk* and *rst* are the main clock and reset signal, while *x_in* serves as the input signal for the entire equalizer; *w*0–*w*8 are the weight update ports of the filter.

3.2.3 Design of Error Control Module

The function is to compare the errors between the input signal and the nondestructive signal, and then send the results to the weight coefficient updating module. We use multiplier and adder to realize the function. The output of the 16 bit signed number after this device and intercept it before it is sent to the update module. The schematic diagram is shown in Fig. 3. The input y is the input signal of the transversal filter and $dout\_mult$ is the output signal of the error control module.

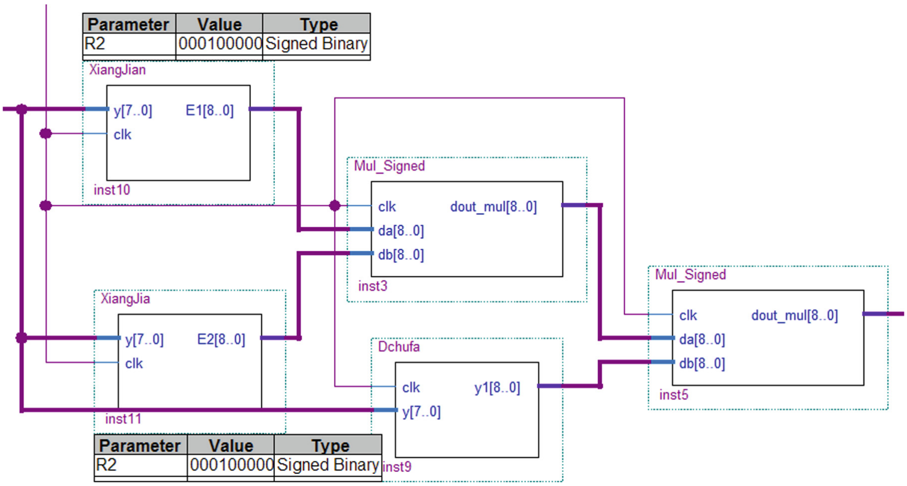


Fig. 3. Error module design

The Xiangjian part is a subtracter, the Xiangjia part is an adder, and the two Mul_Signed modules are multipliers and the Dchufa part is low flip-flop. The Xiangjian part is extended to 9-bit register storage to avoid the signal overflow.

3.2.4 Module Design of Updating Weight Coefficient

The module is used to calculate the filter tap order, which is realized by the calculation of coefficient updating, and then sent it into the filter to update its coefficient.

As shown in Fig. 4, it can be divided into three parts. First, we can multiply the step size by xe to calculate the updated coefficient, then superpose it with the filter coefficient on the previous moment, and then send it into the transversal filter module by $W0\_gengxin \sim W8\_gengxin$. Put the weight coefficient as output and send it to the next module.

In Fig. 4, the Djianwei part is used for data interception, the first eight bits are intercepted as output signals and ignore the lowest order. The Mul_XISHU part is the coefficients multiplication. The d_YANCHI conducts Dtrigger part to delay by one clock, while the gengxin part is used to update the module and add the previous moment coefficients and send it as an output to the filter.

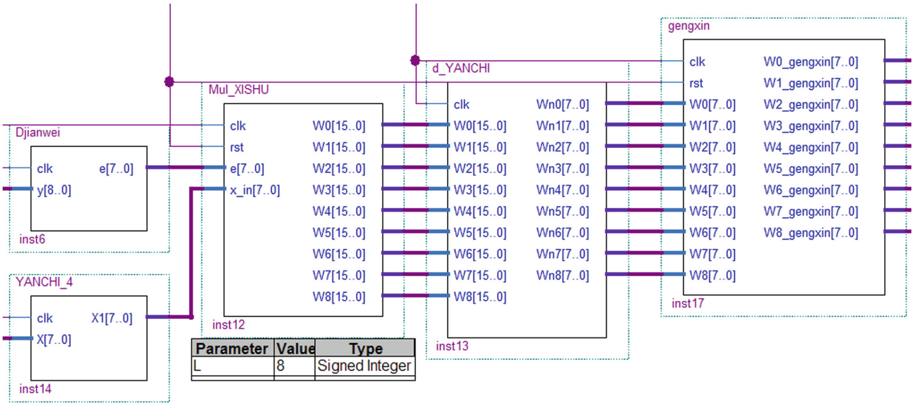


Fig. 4. Design of weight coefficient updating module

4 Debugging and Result Analysis

According to the design scheme, we can further verify the waveform verification, using the FPGA platform to obtain the waveform diagram as follows (Fig. 5).

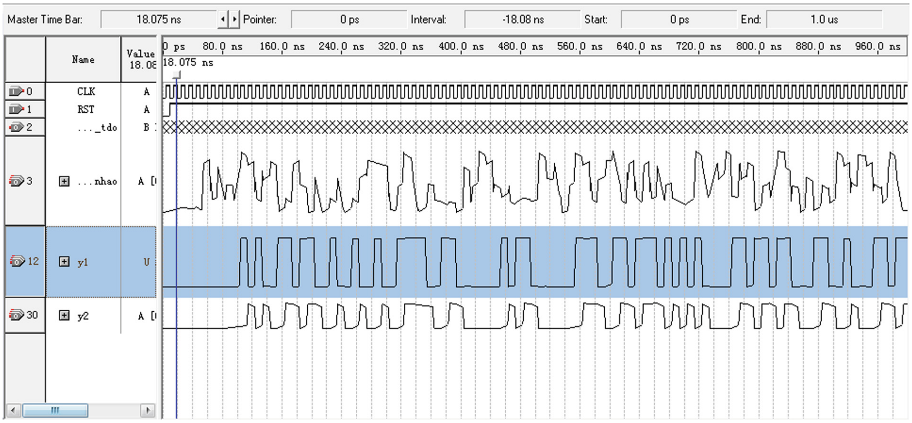


Fig. 5. FPGA simulation waveform diagram

The LMS filter is designed in FPGA. The signal passes through the channel inter-symbol interference and the noise interference at the sequence number 3 in the diagram, and is used as the input signal of the filter. The y1 signal is the output signal filtered by the filter and y2 is the test signal. We can clearly see that the input signal is very difficult to distinguish between the high and low levels, resulting to the error of the decision, and the correct baseband signal cannot be obtained. From the recovered output signal, we can make a very clear decision. It is easy to discriminate because the high and low levels are distinct.

5 Summary

To sum up, the filtered signal of LMS filter designed in FPGA can recover the original baseband signal, which proves the idea of the design. Therefore, we can prove that the design can meet the requirements.

Acknowledgments. This work is supported by Key projects of natural science foundation of Fujian province of China. No.: 2018H01010223, Major projects of Quanzhou science and technology bureau of China. No.: 2017G012, and talent Project of Fujian university of technology No.: GY-Z17011, No.: GY-Z18088.

References

1. Song, N., Alokozai, W.U., de Lamare, R.C.: Adaptive widely linear reduced-rank beamforming based on joint iterative optimization. IEEE Signal Process. Lett. **21**(3), 265–269 (2014)
2. Vorobyov, S.A.: Principles of minimum variance robust adaptive beamforming design. Signal Process. **39**(12), 3264–3277 (2013)
3. Zhang, X.: Modern Signal Processing. Tsinghua University Press, Beijing (1999)
4. Krishnamurthy, V., Logothetis, A.: Adaptive nonlinear filter for narrowband interference suppression in spread spectrum CDMA. IEEE Trans. Signal Process. **44**, 629–639 (1996)
5. Chen, L., Zheng, J.: A broadened and deepened anti-jamming technology for high-dynamic GNSS array receivers. IEICE Trans. Commun. **E99-B**(9), 2055–2061 (2016)
6. Zhu, Z., Gao, X., Cao, L., Pan, D., Cai, Y., Zhu, Y.: Analysis on the adaptive filter based on LMS algorithm. Optik – Int. J. Light Electron Opt. **127**, 4698–4704 (2016)
7. Fuchs, J.: Identification of real sinusoids in noise, the global matched filter approach. In: Proceedings of the 15th IFAC Symposium on System Identification, pp. 1127–1132 (2009)
8. Mustafa, R.U., Nawaz, M.S., Ferzund, J.: Early detection of controversial urdu speeches from social media. Data Sci. Pattern Recogn. **1**(2), 26–42 (2017)
9. Pan, J.-S., Kong, L.P., Sung, T.-W.: Hierarchical routing strategy for wireless sensor network. J. Inf. Hiding Multimed. Signal Process. **9**(1), 256–264 (2018)
10. Meng, S.-H., Hu, S.-B., Chui, H.-C.: A broadband ASE light source-based FTTX RoF-WDM optical network system. J. Netw. Intell. **2**(1), 162–170 (2017)

Fast Clustering Algorithm for Construction Areas Based on Spatiotemporal Trajectory of Engineering Vehicles Group

Siqi Gao[1,2], Fumin Zou[1,2](✉), Lyuchao Liao[1,2], Meirun Zhang[1,2], Yixi Peng[3], and Subing Song[3]

[1] Beidou Navigation and Smart Traffic Innovation Center of Fujian Province, Fujian University of Technology, Fuzhou 350118, Fujian, China
gsq281221@gmail.com, fjachao@gmail.com, fmzou@fjut.edu.cn, laoniu8706@126.com

[2] Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China

[3] Fujian Provincial Traffic Information Communication and Emergency Disposal Center, 28F, Traffic Complex, No. 18 Dongshui Road, Fuzhou 350118, Fujian, China
pengyixi123@163.com, subin.song@foxmail.com

Abstract. Spatiotemporal trajectory data mining of vehicle groups is an important approach to identify the travel mode of mobile objects. A time series based fast density peaks clustering (TSDPC) algorithm aiming at rapid discovery of construction areas with respect to structural knowledge contained in the trajectory of engineering vehicles is proposed. Firstly, high-density areas extraction of engineering vehicles trajectory data characterized the low-frequency is accomplished by partition clustering to establish a cluster candidate set; Then the DPC algorithm based on the improved local density is utilized to conduct the parallel calculation of the candidate set in the region; Finally, the local clustering results are merged to complete global rapid clustering. The field experiment results demonstrate that the TSDPC algorithm improves the clustering efficiency of the spatiotemporal trajectory data of large-scale engineering vehicles group and effectively realizes the rapid discovery of construction areas.

Keywords: TSDPC · Construction areas discovery · Spatiotemporal trajectory Engineering vehicles · Partition clustering

1 Introduction

Infrastructure construction reflects the economic activities of the society and is also related to the good life of the people. With the vigorous development of China's economy, various types of construction areas are innumerable, resulting in rapid changes of urban road network and spatial structure. Geospatial information updates like buildings and road networks are conventionally mainly carried out in-the-field by the relevant departments, which requires high cost and long period. Therefore, it is the

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 33–41, 2019.
https://doi.org/10.1007/978-3-030-04585-2_5

fact that a low-cost and fast discovery method for the construction areas is urgently needed.

Literature [1] applies high-resolution satellite imagery to monitor the actual progress in the project. However, this method featured high cost and information lag as well is difficult to extract information in that remote sensing images are mixed with a variety of features and are easily interfered by obstacles such as clouds. Literatures [2, 3] visualize the progress of infrastructure construction based on aerial imagery of drones and accurately extract markers such as tower cranes in the image to achieve identification of construction areas, which only satisfy with identification for small-scale regions.

With the rapid development and popularization of floating car technology, relevant departments have accumulated a large amount of vehicle trajectory data. Vehicle groups travel patterns and behavior analysis [4] have gradually become research hotspots owing to the trajectory data implies various structured knowledge such as traffic geography patterns and user driving behavior habits [5]. However, it proves that the recent research on regional discovery based on the analysis of taxi trajectory data [6–8] fails to identify the construction areas.

As a result, the engineering vehicles group whose trajectory data contains structural knowledge of construction areas is taken as the research object in this paper, and the rapid discovery of construction areas is implemented by clustering analysis of its spatiotemporal trajectory.

2 Problem Statement and Related Concepts

The traffic trajectory data is spatial information collected by the vehicle satellite positioning terminal. Engineering vehicles usually travel to and from roads, construction areas, and loading yards in account of their functionality, which the trajectory is relatively unified and the duration in the construction areas is relatively dense, resulting in a spatiotemporally obvious gathering situation [9].

Definition 1 (Traffic trajectory sequence): Given an engineering vehicle, trajectory sequence consisting of space, time and other information collected periodically during driving is denoted as:

$$\mathrm{S}_i = \langle P_0(id_0, x_0, y_0, t_0),\ P_2(id_1, x_1, y_1, t_1), \ldots\ldots,\ P_n(id_n, x_n, y_n, t_n)\rangle \tag{1}$$

Where $P_n(id_n, x_n, y_n, t_n)$ indicates spatiotemporal data consisting of vehicle ID, longitude, latitude, and time information, $\mathrm{n} = \mathrm{len}(\mathrm{S}_i)$ is length of the traffic trajectory sequence, namely the number of points. Figure 1 shows the 3D trajectory of four engineering vehicles, all of which have a distinct gathering situation.

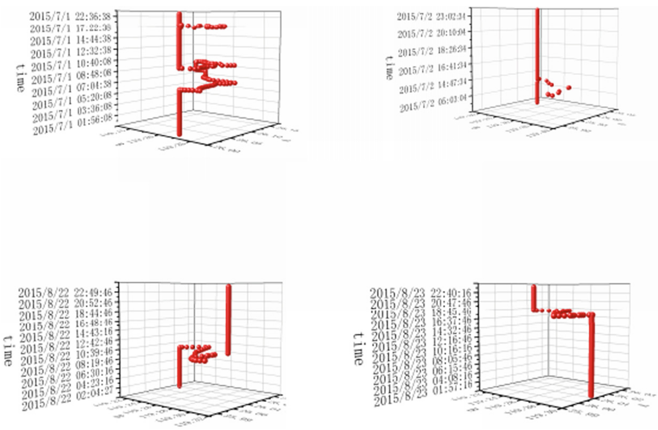


Fig. 1. 3D trajectory graph of engineering vehicles

Definition 2 (Grid partition): Suppose a geospatial space S characterized latitude and longitude range: $lon_{min} \sim lon_{max}, lat_{min} \sim lat_{max}$, we divide S into M*N space grids:

$$\begin{cases} \mathrm{M} = (lon_{max} - lon_{min})/l \\ \mathrm{N} = (lat_{max} - lat_{min})/h \end{cases} \tag{2}$$

For any vehicle in the S, its track point (lon,lat) can be mapped into the corresponding grid:

$$\begin{cases} \mathrm{m} = (lon - lon_{min})/l \\ \mathrm{n} = (lat - lat_{min})/h \end{cases} \tag{3}$$

Where l *and* h represent grid sizes.

The spatial grid set $\mathrm{G} = (g_i)$ can be obtained from formula (2), for arbitrary i, $\mathrm{j}(i \neq j)$, $g_i \cap g_j = \emptyset$.

Definition 3 (Local density ρ_i [10]): The local density of the sample i is the number of samples in the neighborhood of the cut-off distance d_c of the sample i, its local density ρ_i can be defined as follows:

$$\rho_i = \sum\nolimits_{j \neq i} \chi(d_{ij} - d_c)\ (j = 1, 2, \ldots, n) \tag{4}$$

Where $\chi(\mathrm{x}) = 1$, if $\mathrm{x} < 0$ or $\chi(\mathrm{x}) = 0$, otherwise, d_c is specified by the user, d_{ij} is the Euclidean distance between the samples i and j.

Definition 4 (the distance δ_i of the sample i [10]): δ_i is the minimum distance between the point i and any other point with higher local density, which can be calculated as follows:

$$\delta_i = \min_{j:\rho_{j} \succ \rho_i}\left(d_{ij}\right) \tag{5}$$

For the point with the highest density, we define $max_j\left(d_{ij}\right)$ as its δ_i.

3 Algorithm Analysis and Design

The DPC algorithm is based on two features of an ideal clustering center [10]: (1) the local density ρ_i of the cluster center is large, namely the density of its neighbors is not more than itself; (2) The distance δ_i of different cluster center points is relatively far. By calculating these two parameters, the two-dimensional decision graph is drawn to select the cluster center by $\gamma = \rho * \delta$. And then it compares the local density with the cluster boundary density and marks the cluster members as the core and cluster halo, completing one-step allocation strategy for remaining samples.

According to the formula (3), the cut-off distance d_c is a discrete value. Therefore, when it is used to calculate the local density, the probability of conflicting on a dataset with uniform distribution is relatively large, that is, different data points have the same local density value. Moreover, the calculation needs to traverse the entire dataset, and the time complexity of the algorithm is approximately $o(n^2)$, which it is not practical when the dataset is large. Therefore, this paper proposes a fast density peaks clustering TSDPC algorithm in terms of spatiotemporal distribution characteristics of engineering vehicles trajectory.

Definition 5 (TS local density $\rho_i^{'}$): Set the time window sequence TS(i) to be a set of track points within time window range of the sample i, $\rho_i^{'}$ is defined as follows:

$$\rho_i' = \sum\nolimits_{j \in TS(i)} exp\left(-d_{ij}^2\right) \tag{6}$$

While the engineering vehicles working in the construction areas, the spatiotemporal trajectory of which is highly aggregated. Therefore, it can be assumed that the local density ρ_i of the sample point is proportional to its local density $\rho_i^{'}$ within TS(i) (Table 1).

Definition 6 (TS outliers): Given $N_{dist}(i) as$ the boundary distance of the N points in each sample's TS and Threshold as the outlier threshold, and the set Outlier of the sample o whose $N_{dist}(o)$ is greater than Threshold is an outlier set:

$$N_{dist}(i) = \max_{j \in TS(i)}\left(d_{ij}\right) \tag{7}$$

$$\text{Threshold } = \frac{1}{N}\sum\nolimits_{i=1}^{N} N_{dist}(i) \tag{8}$$

$$\text{Outlier} = \{o | N_{dist}(o) \succ \text{Threshold}\} \tag{9}$$

Table 1. Overall framework of the TSDPC algorithm

```
TSDPC Algorithm
```

```
Intput:
D:dataset
l、h:width of grid measured in degree
ε: density threshold of grid
TS: the window for searching
Output:
C: a set of clusters
Method:
Module1:/* grid-based partitioning phase*/
/*map the point into the Grid*/
Step1:Execute formula (1) to split S into grids
Step2:create an empty dict_G using the matching re-
sults
Step3:for each point in D do
Step4:map each point into correspond g∈dict_G
/*count points for each grid to update index*/
Step5:if g_ID in dict_G then
dict_G[g_ID].append(currentpoint)
Step6:else create an new dictionary item [g_ID: cur-
rentpoint]
Step7:End if
Step8:end for
Step9:create an empty dict_C
Step10:For each g_ID in dict_G:
dict_C[g_ID]←len(dict_G[g_ID])
/*select the high density Grid*/
Step11:create a set G
Step12:for each g_ID in dict_C
Step13:if dict_C[g_ID]> ε then add dict_G.key and
dict_G.values to G
Step14:end if
Step15:end for
Generate a PartitionedPointsRDD based assigned points
named GRDD
Module2:/* Partition clustering implementation by
TSDPC algorithm*/
Step1:for each grid in GRDD do
Step2:execute TSDPC algorithm
/*parallel computing with TSDPC algorithm*/
Step3:for each point in grid
Step4:calculate ρi using formula 5
Step5:calculate δi using formula 4
Step6:end for
Step7:end for
Step8:calculate γi = ρi * δi to generate local clusters
Step9:delect outliers by formula(6)(7)(8)
/*Merging phase*/
Step10: merge local clusters to establish final clus-
tered C
```

4 Experimental Design and Analysis of Results

4.1 Dataset of Engineering Vehicles

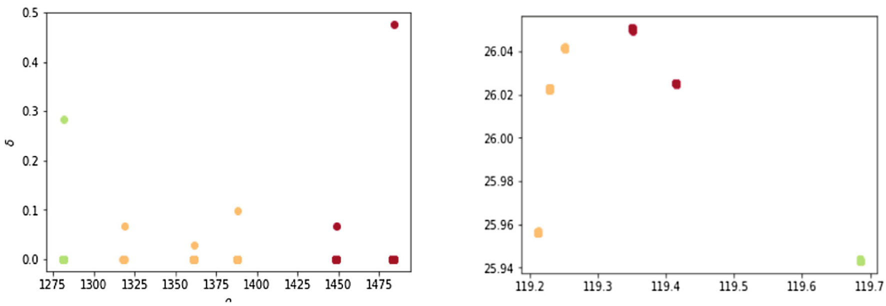
This section intents to verify the TSDPC algorithm on the real dataset. The experimental data comes from the spatiotemporal trajectory dataset of 40 engineering vehicles in July 2015 in Fuzhou City supported by Beidou Navigation and Smart Traffic Innovation Center of Fujian Province (The latitude and longitude range are 119.2032° ~ 119.7356°, 25.9114° ~ 26.1611° respectively, and the sampling interval is 15 ~ 30 s).

4.2 Data Preprocessing

The first is to convert trajectory time to Unix time and sort them, eliminating points with the same time. The trajectory data is then projected as map data with a reference coordinate system of WGS1984. For the maximum width of the main road in Fuzhou City is approximate 40 meters, taking into account the entire calculation scale and accuracy, we set the fixed step size of the grid side length § = 100 meters, corresponding to the accuracy of the satellite positioning data of Fuzhou City 0.001°, that is, $l = h = 0.001°$ in formula (1-2). And m and n are the row and column index respectively. The trajectory points are ultimately mapped into corresponding spatial grids.

4.3 Experimental Environment and Results Analysis

We conduct the experiment on a desktop computer with a core i7 3.6 GHz processor, 64-bit Windows10 operating system, 8G RAM and 500G hard disk by running Python3.6. The trajectory points of engineering vehicles are mapped to the set G, and the high-density grids are extracted as a cluster candidate set. In this case, due to the data set is large and is closely distributed, only about 60% of the dense grids remained. The TSDPC algorithm executes the formula (5) to calculate the local density and determine the cluster centers by $\gamma = \rho * \delta$. The specific clustering results are shown in Fig. 2.



(a) Decision graph (b) Clustering result graph

Fig. 2. Clustering results of TSDPC algorithm

The efficiency experiment comparing the time required for the DPC and the TSDPC to process the massive dataset of the engineering vehicles is divided into four groups: 100,000, 1 million, 5 million, and 10 million track points from the dataset. The result is shown in Fig. 3.

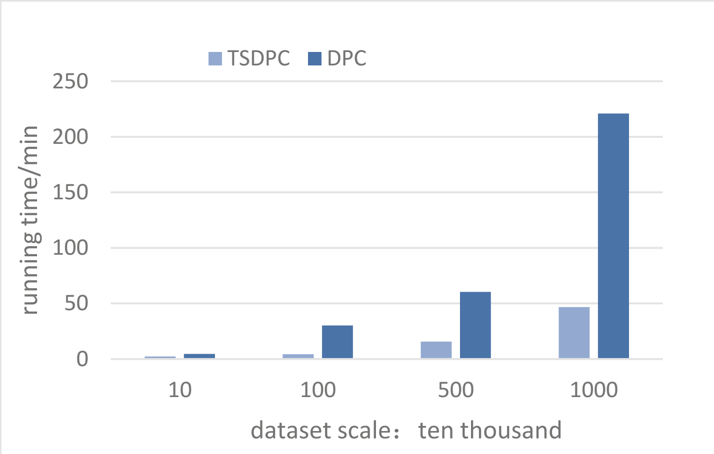


Fig. 3. Algorithm efficiency comparison chart

The original PDC algorithm has short running time on the small-scale dataset, but the running time of which increases exponentially as the scale does since it cannot segment the data set. The TSDPC algorithm takes a small amount of time to split the dataset, so its efficiency on the small-scale dataset is equivalent to that of the DPC. However, when the dataset increases, its running time is relatively flat and the algorithm is more scalable.

We further validate the construction areas by remote sensing satellite image and in-the-field verification. Figure 4 shows two found construction areas. Figure 4 (a) is the remote sensing satellite image before construction, (b) is after construction, where the new building is found in the comparison chart (blue circle in the red box), and (c) (d) is before and after the construction of Huxiu Road.

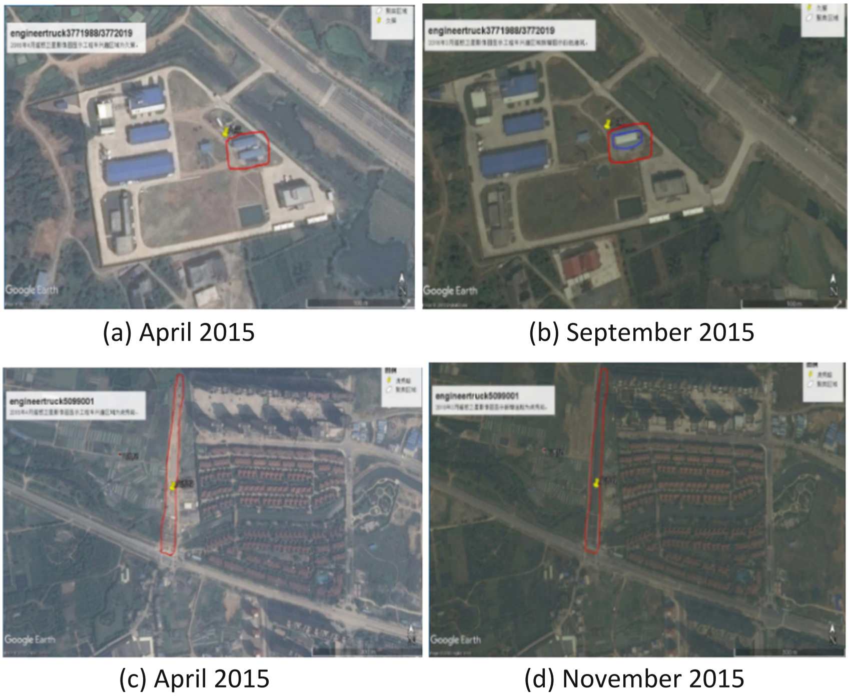


Fig. 4. Remote sensing satellite comparison images

5 Conclusion

Aiming at the need for rapid discovery of construction areas, a fast clustering algorithm based on spatiotemporal trajectory data of engineering vehicles group is proposed. Spatial grid is introduced to address the calculation drawback that the original DPC algorithm is not applicable to large-scale dataset, which achieves the parallel processing of the algorithm. Finally, the verification combining remote sensing satellite imagery with in-the-field identification reveals that the method can quickly and effectively discover the construction areas.

Acknowledgments. We would like to appreciate the constructive comments and financial supports from projects of Model analysis, visualization and decision optimization support for urban passenger flow data (No. 41471333), Smart passenger vehicle positioning terminal: [Grant number: GY-H-17040], Development and Demonstration Application of Fujian New Energy Vehicle Supervision Service System: [Grant number: GY-Z160141), RongJingXin Technology (No. 2017-1034). Beidou Navigation and Smart Traffic Innovation Center of Fujian Province is also acknowledged for supporting the experimental dataset.

References

1. Behnam, A., Wickramasinghe, D.C., Ghaffar, M.A.A., et al.: Automated progress monitoring system for linear infrastructure projects using satellite remote sensing. Autom. Constr. **68**, 114–127 (2016)
2. Ham, Y., Han, K.K., Lin, J.J., et al.: Visual monitoring of civil infrastructure systems via camera-equipped Unmanned Aerial Vehicles (UAVs): a review of related works. Vis. Eng. **4**(1), 1 (2016)
3. Yu, B., Niu, W., Wang, L., et al.: A tower crane extraction method based on multi-scale adaptive morphology. Remote Sens. Technol. Appl. **28**(2), 240–244 (2013)
4. Zhang, Y., Guo, S., Xing, K., et al.: Mining algorithm of spatial-temporal co-occurrence pattern based on vehicle GPS trajectory. In: IEEE International Conference on Signal and Image Processing, pp. 142–145. IEEE (2017)
5. Liao, L.C., Jiang, X.H., Zou, F.M.: A spectral clustering method for big trajectory data mining with latent semantic correlation. Acta Electron. Sin. **43**(5), 956–964 (2015)
6. Yuan, J., Zheng, Y., Xie, X.: Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 186–194. ACM (2012)
7. Liu, S., Wang, S.: Trajectory community discovery and recommendation by multi-source diffusion modeling. IEEE Trans. Knowl. Data Eng. **29**(4), 898–911 (2017)
8. Garg, N., Ramadurai, G., Ranu, S.: Mining bus stops from raw GPS data of bus trajectories. In: 2018 10th International Conference on Communication Systems & Networks (COMSNETS), pp. 583–588. IEEE (2018)
9. Liao, L., Wu, J., Zou, F., et al.: Trajectory topic modelling to characterize driving behaviors with GPS-based trajectory data. J Internet Technol. **19**(3), 815–824 (2018)
10. Rodriguez, A., Laio, A.: Clustering by fast search and find of density peaks. Science **344** (6191), 1492 (2014)

The Development and Analysis of Electromagnetic Compatibility for Vehicles

Yi-Nung Chung[1(✉)], Ming-Sung Chiu[1], Chao-Hsing Hsu[2], and Kai-Chih Chuang[1]

[1] Department of Electrical Engineering, National Changhua University of Education, Changhua 500, Taiwan
ynchung@cc.ncue.edu.tw, Hugojely@gmail.com, kisime2002@yahoo.com.tw

[2] Department of Information and Network Communications, Chienkuo Technology University, Changhua 500, Taiwan
hsu@cc.ctu.edu.tw

Abstract. There are more and more electronic products in the vehicle, such as communication device, navigation system, audio, video equipment, and etc. Because the electronic devices of vehicle design are very complicated, therefore it will have interference between the equipment. To avoid the interference of each other is very important. It means that the vehicle with electromagnetic compatibility design is necessary. The design needs to follow the electromagnetic compatibility regulations, through its electromagnetic interference and electromagnetic tolerance specifications and other test items. In this paper, the design of electromagnetic compatibility for the vehicle to avoid the electromagnetic interference is investigated. Based on the test, this design has better electromagnetic tolerance, lower cost, and can achieve high quality electronic products.

Keywords: Electromagnetic compatibility · Electromagnetic interference Electromagnetic tolerance

1 Introduction

There are more and more electronic products in the vehicle. The problem of electromagnetic interference (EMI) [1] is more serious day by day. If there is not electromagnetic compatibility (EMC) [2–4] development, it will have some effects. So, the electromagnetic compatibility design is necessary for the vehicle. The effects include that the audio screen may flash, the driving recorder may be reopened, the digital signal may be disturbed, and etc. These functional failures may not affect the driving safety, but will cause customer's bad feeling. If the safety components are disturbed, for example the automatic brake judgment system is abnormal, which may cause the car accident and to affect the driver's life. Once the failure occurs, the safety of the driver and the passenger may be endangered. Therefore, in recent years, the international vehicle manufactory has begun to require interior safety-related electronic products to pass ISO 26262 and to ensure the safety [5, 6]. Especially, for the next generation smart automobile industry, the electromagnetic compatibility technology is more important.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 42–50, 2019.
https://doi.org/10.1007/978-3-030-04585-2_6

The development of EMC for electronic products should be conducted at the beginning. For example, the product sample has been designed but the EMC test is unable to pass the specification. Then, the EMC problem is not easy to solve. Usually, the design of circuit and mechanism cannot make great changed and the cost of changing design is huge also. If there is this kind of problem, the layout of printed circuit board (PCB) should be developed again and then make the EMC test again also. Because the mechanism mold has been made, so it cannot make lot of changes. The production process might not be typically. However, because these are not typical production process, the assembly and contact may not well. Sometimes this kind of product can pass the EMC in self-test, but the EMI problem still existed and cannot pass EMC test when the vehicle is shipped to the customer. It will cause the customer to feel bad and need to spend other test cost. Therefore, we need to conduct a feasibility evaluation based on the EMC test specification which is provided by the customer at the beginning. The necessary EMC design needs to be added to the circuit, mechanism, and wires. And the EMC test plan needs to be provided at the beginning of the design [5–8]. In this paper, the development and analysis of EMC for vehicles is investigated.

2 System Design

EMC related design and protection can be divided to four parts, which are circuit design, PCB design, mechanism design, and wire design. The circuit design is the first step for product design. The interference sources include oscillator on the circuit, high speed signal, and power circuit. In order to reduce the interference of electromagnetic radiation, it uses the resistor in the circuit to low down the current and power. It needs to use the filters to filter out specific frequency signal and to stabilize the power supply. The circuit design also needs to prevent the externally interference for the system, such as the surge attack, lighting strike, and unstable power situation.

After the circuit design is completed, the circuit layout of the PCB is the next step. At the beginning, the position of the parts in PCB is very important. For example, the analog circuit must be isolated from the digital signal, the easy interfered components are placed in the center, high-speed signal component cannot put at the edge, and etc. The design must consider to prevent the static electricity from affecting the electronic parts in the PCB. The best connection is the whole circle continuously. Moreover, it also needs to consider the placement of the electronic countermeasure components, such as the ESD protection parts are close to the interference source and the shortest path is connected to the power source. Let the interference signal do not affect the central sensitive component.

Inside the electronic product, there is a mechanism body to connect different functions of the PCB board, such as lens, liquid crystal display, and etc. At the beginning of design, it is necessary to understand the material and appearance of product. The material will be related to the direction of the EMC design. If it is a plastic case, the EMC design will be different, because the plastic case is useless for the internal EMI radiation and external interference. For the static protection must consider the gap between the internal electronic parts to ensure the non-discharge safety. If the space is not enough to reserve a non-discharge safety distance, then must consider internal parts' position. We need to leave a ground point, let the static electricity discharge safely.

The vehicle electronic product has a power signal line connected to the power supply or signal input device of the vehicle. The connection cable may bring out the interference source on the electronic component. And the connection line may become the transmitting antenna which will affect other parts. Moreover, the connection line may also become the receiving antenna to affect the circuit. There are two kinds of connection cable which are the single layer and double layers. The single layer cable uses metal to woven mesh. The double layer cable is added the aluminum foil outside of metal mesh. The EMI can be reduced if this kind of cable is used. The wire material can also be added with the element of the iron powder core material, and the wire can pass through the component to achieve the filtering effect. By using this kind of design, it can reduce the external signal interference and to obtain the better EMC effect.

In this study, the EMC circuit designs are developed. In order to reduce the internal EMI for image signal transferring and enhance the EMI protection capability, one low pass filter circuit is developed. The circuit design of low pass filter is shown in Fig. 1. The static electricity protection circuit is applied to the static electricity discharge, which is shown in Fig. 2. In order to avoid the reverse voltage to attack the system, one reverse voltage protection circuit shown as Fig. 3 is developed. The reset protection circuit shown in Fig. 4 is designed for resetting system when the unusual shot down is happened.

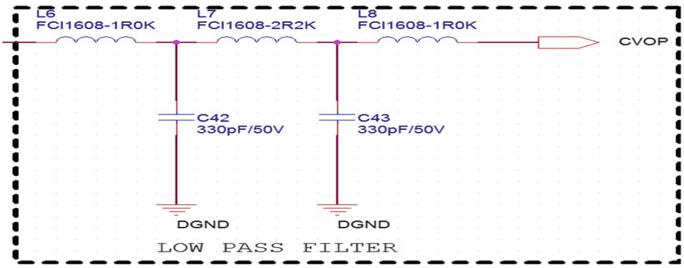


Fig. 1. The design of low pass filer

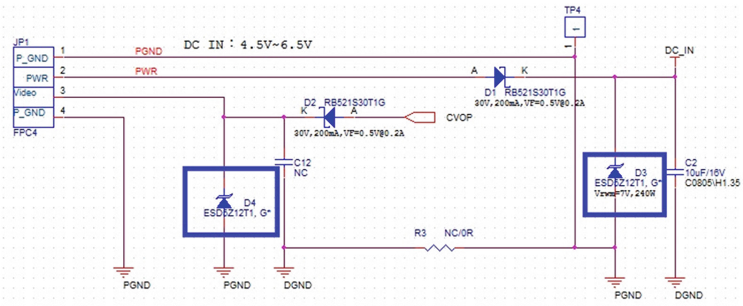


Fig. 2. The static electricity protection circuit

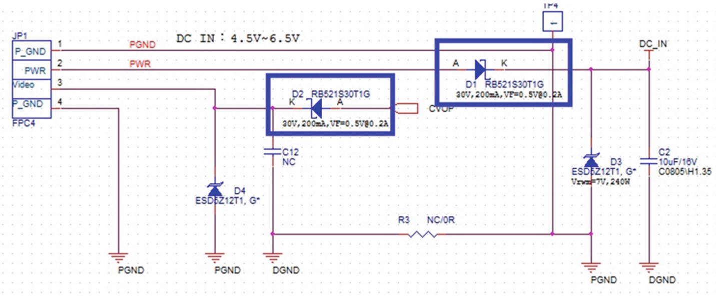


Fig. 3. The reverse voltage protection circuit

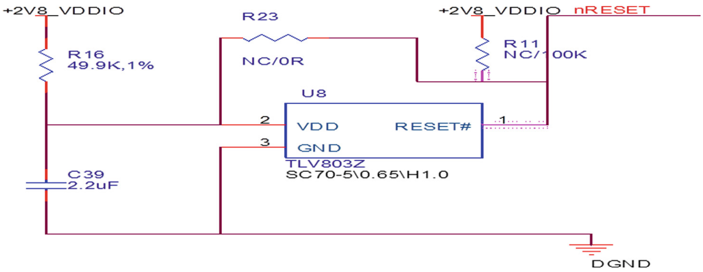


Fig. 4. Reset protection circuit

3 Experiment and Analysis

The vehicle guide image assisting system is applied in this experiment. The block diagram of the circuit structure is shown in Fig. 5. The connector has four wires which include power source, power ground point, image signal, and image ground point. The power source of the sensor is 5 V DC supplied by the vehicle, which requires to convert

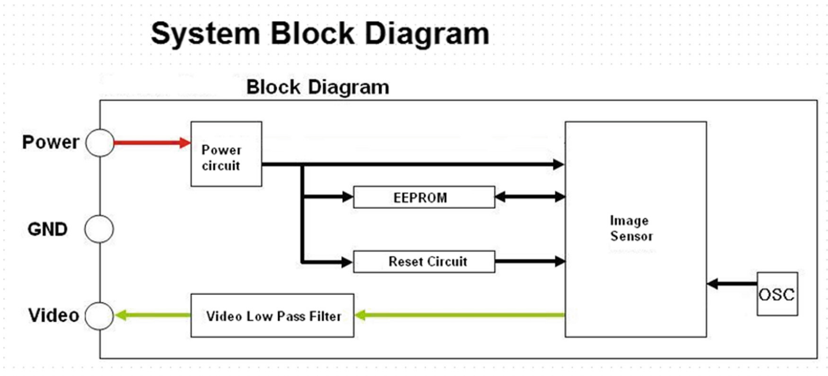


Fig. 5. Block diagram of the circuit structure

by the battery. The EEPROM is used to record the parameters of the sensor. In this system, the image output is passed through a low-pass filter to reduce the EMI radiation interference value. The test includes test specifications, limit standards, compliance levels, power supply, component location, and wire length. The conducted interference measurement configuration diagram is shown in Fig. 6. High current injection withstand test configuration diagram is shown in Fig. 7. Radiation tolerance test configuration diagram is shown in Fig. 8. Radiation interference test configuration diagram is shown in Fig. 9. Static resistance test configuration diagram is shown in Fig. 10. These test will follow the specific standards provided by customer.

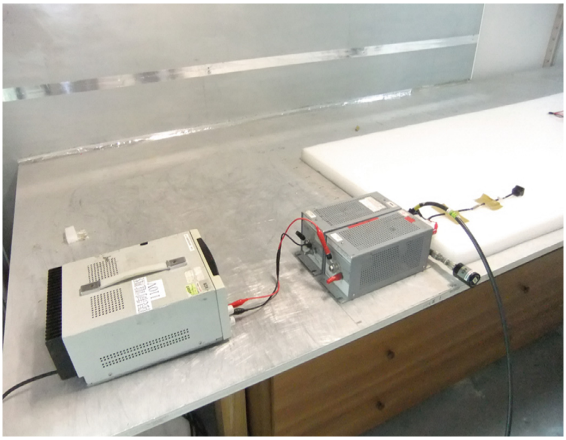


Fig. 6. Conducted interference measurement configuration diagram

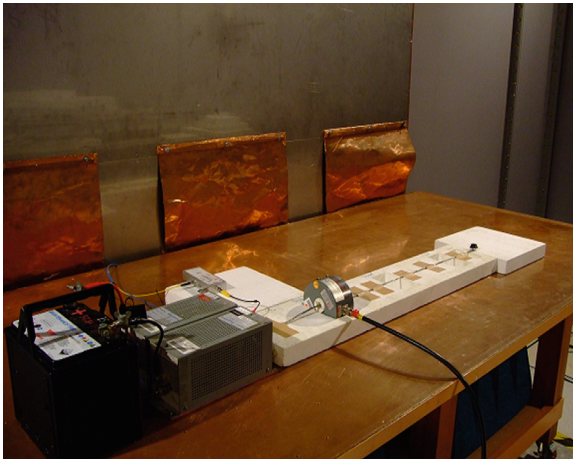


Fig. 7. High current injection withstand test configuration diagram

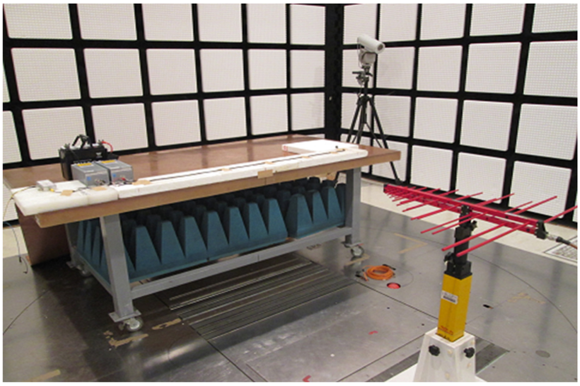


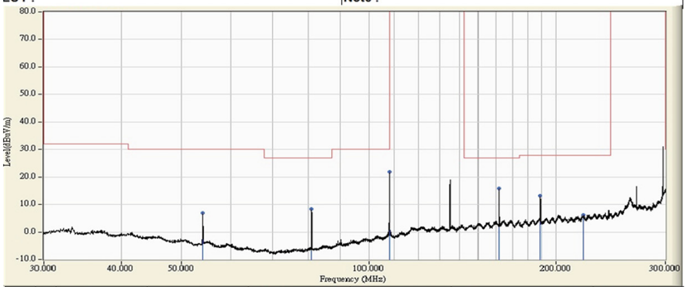
Fig. 8. Radiation tolerance test configuration diagram

Fig. 9. Radiation interference test configuration diagram

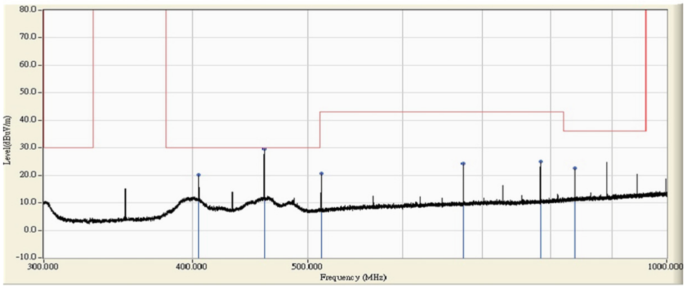
Fig. 10. Static resistance test configuration diagram

Based on the first time of the EMC test results, three items are failed. (1) The BCI high current tolerance without using the three-layer isolation line is disturbed. (2) The RE radiation interference did not add isolation network wire, the oscillator R is zero ohm, and C is open circuit. Based on the test, it is found that the interference is occurred by main frequency 27 MHz signal. The radiation interference measurement value of first time test is shown in Fig. 11. (3) The static test does not leave the protective ground around the PCB. So when the air discharge at 25 kV will hit the electronic component.

Limit : CISPR25_2008_Class3_01M_AV	Margin : 0
Probe : Biconical_0830 - HORIZONTAL	Power : DC 5V
EUT :	Note :



		Frequency (MHz)	Correct Factor (dB)	Reading Level (dBuV)	Measure Level (dBuV/m)	Margin (dB)	Limit (dBuV/m)	Detector Type
1		54.000	-39.624	46.565	6.941	-23.059	30.000	AVERAGE
2		81.000	-42.114	50.380	8.266	-18.734	27.000	AVERAGE
3	*	108.000	-37.272	59.207	21.935	-8.065	30.000	AVERAGE
4		162.000	-32.510	48.365	15.855	-11.145	27.000	AVERAGE
5		189.000	-31.132	44.245	13.113	-14.887	28.000	AVERAGE
6		221.200	-29.904	36.008	6.103	-21.897	28.000	AVERAGE

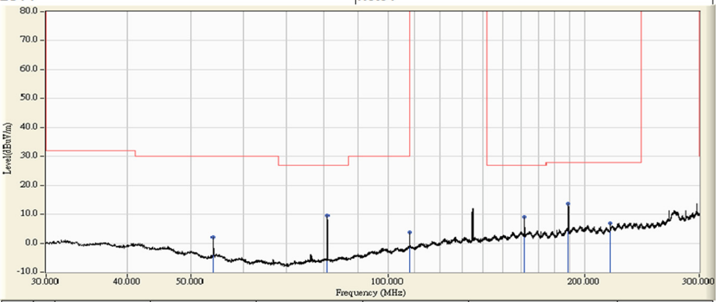


		Frequency (MHz)	Correct Factor (dB)	Reading Level (dBuV)	Measure Level (dBuV/m)	Margin (dB)	Limit (dBuV/m)	Detector Type
1		405.000	-30.068	50.112	20.044	-9.956	30.000	AVERAGE
2	*	459.000	-29.206	58.802	29.596	-0.404	30.000	AVERAGE
3		513.000	-28.351	48.888	20.537	-22.463	43.000	AVERAGE
4		675.000	-26.187	50.435	24.248	-18.752	43.000	AVERAGE
5		783.000	-25.328	50.306	24.978	-18.022	43.000	AVERAGE
6		837.000	-24.750	47.284	22.534	-13.466	36.000	AVERAGE

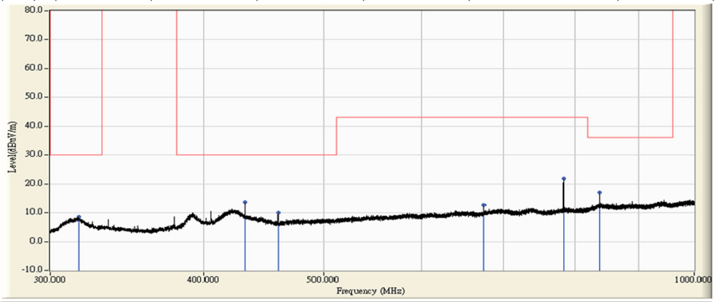
Fig. 11. The first time radiation interference measurement value

After the EMC design proposed in this study. The test results can pass the required regulation. (1) First, the three layers of isolation wires are used, so the BCI test can meet the requirement of level A. (2) We add 33R and 33P at the 27 MHz oscillator, so it can pass the RE test. After using three layers of isolation wires, the radiation energy reduction obviously. The second time test result is shown in Fig. 12. (3) We add the surrounding protective ground to let the air discharge 25 kV hit the ground and to vent the energy. So it can pass the PCB electrostatic test. According to EMC design guidelines, it can reduce the number of PCB revisions and test cost. It also can achieve stable electromagnetic compatibility effect.

Limit : CISPR25_2008_Class3_01M_AV	Margin : 0
Probe : Biconical_0830 - HORIZONTAL	Power : DC 5V
EUT :	Note :



		Frequency (MHz)	Correct Factor (dB)	Reading Level (dBuV)	Measure Level (dBuV/m)	Margin (dB)	Limit (dBuV/m)	Detector Type
1		54.000	-39.624	41.635	2.011	-27.989	30.000	AVERAGE
2		81.000	-42.114	51.620	9.506	-17.494	27.000	AVERAGE
3		108.000	-37.272	40.996	3.724	-26.276	30.000	AVERAGE
4		162.000	-32.510	41.578	9.068	-17.932	27.000	AVERAGE
5	*	189.000	-31.132	44.819	13.687	-14.313	28.000	AVERAGE
6		219.600	-29.996	36.777	6.781	-21.219	28.000	AVERAGE



		Frequency (MHz)	Correct Factor (dB)	Reading Level (dBuV)	Measure Level (dBuV/m)	Margin (dB)	Limit (dBuV/m)	Detector Type
1		316.200	-32.081	40.685	8.604	-21.396	30.000	AVERAGE
2	*	432.000	-29.728	43.479	13.751	-16.249	30.000	AVERAGE
3		459.000	-29.270	39.411	10.141	-19.859	30.000	AVERAGE
4		675.000	-25.837	38.496	12.659	-30.341	43.000	AVERAGE
5		783.000	-25.060	46.942	21.882	-21.118	43.000	AVERAGE
6		837.000	-24.069	41.129	17.060	-18.940	36.000	AVERAGE

Fig. 12. The second time radiation interference measurement value

4 Conclusion

Because there are more and more electronic products in the vehicle, so the EMI problem in the system is more serious. In this paper, one effective EMC methodology is developed. According to the experimental results, all the test items can successfully pass the required conditions by using this method. It can effectively reduce the number of PCB revisions, reduce EMC countermeasures, test time, and costs. Therefore, the stable electromagnetic compatibility products can be obtained. Based on the analysis, if the EMC design can be conducted in the early stage, the circuit development and product process can be easier. Based on the experimental results, the EMC algorithm proposed in this paper is effective and useful.

Acknowledgments. This work was supported by the Ministry of Science and Technology under Grant MOST 106-2221-E-018-028.

References

1. ISO 7637-3: Road vehicles – electromagnetic disturbances from conduction and coupling – part 3: electrical transient transmission by capacitive and inductive coupling via lines other than supply lines (2007)
2. Brüns, H.-D., et al.: Modeling challenging EMC problems. IEEE Electromagn. Compat. Mag. **6**(3), 45–54 (2017)
3. Bonci, Alessio, Boni, Enrico, Carobbi, Carlo F.M.: A compact, broadband, calculable electromagnetic field source for quality assurance in EMC testing. IEEE Trans. Electromagn. Compat. **60**(5), 1215–1222 (2018)
4. Matsushima, Sayaka, Matsushima, Tohlu, Hisakado, Takashi, Wada, Osami: Trends of EMC standards for automotive network devices and communication quality of ether net in relation to parameters of pulse disturbances. IEEE Electromagn. Compat. Mag. **7**(1), 46–50 (2018)
5. ISO 11452-4: Road vehicles – component test methods for electrical disturbances from narrowband radiated electromagnetic energy – part 4: harness excitation methods (2011)
6. ISO 7637-2: Road vehicles – electrical disturbances from conduction and coupling – Part 2: electrical transient conduction along supply lines only (2011)
7. Hung, M.-H., Chen, Q., Chen, Y.: A review of handling continuous and unknown attribute values of C4.5 algorithm. J. Netw. Intell. **3**(1) (2018)
8. Hsu, S.-P., Chung, Y.-N., Hsu, Y.-C.: Applying power management and wireless sensing system to plant factory. J. Inf. Hiding Multimed. Sign. Process. **7**(2), 419–424 (2016)

Non-linear Random Change Differential Evolution for Multi-objective Resource Allocation Problem

Jierui Liu[1,3], Lyuchao Liao[1,3](✉), Jengshyang Pan[2,3], Fumin Zou[1,3], Guoqian Wang[4], and Qiqin Cai[1,3]

[1] Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China
jieruil204@gmail.com, fjachao@gmail.com

[2] Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, Fujian, China
jengshyangpan@fjut.edu.cn

[3] Fujian Provincial Big Data Research Institute of Intelligent Transportation, Fujian University of Technology, Fuzhou 350118, Fujian, China

[4] Fujian Transport Information and Telecommunication Center, Fujian Communication Department, Fuzhou 350001, Fujian, China

Abstract. The resource allocation problem (RAP) has a wide range of applications in the transportation engineering. The differential evolution algorithm proposed to solve Chebyshev's inequality is essentially an optimization algorithm for the most optimal solution. However, the basic differential evolution algorithm tends to fall into the local optimal solution, resulting in premature convergence of the population, especially in multi-objective resource allocation of transportation engineering. To address these problems, a Non-linear Random Change Differential Evolution was proposed to solve multi-objective resource allocation problems. With tests to the benchmark functions in CEC, the results show that both the non-linear processing and the random changing after the mutation make the result closer to the optimal value. The method make it possible to solve the NP problem of multi-objective allocation in transportation engineering.

Keywords: Multi-objective resource allocation · Differential evolution
Non-linear random change · Premature convergence solution

1 Introduction

People in various regions may have demand for certain kinds of resources, but the resources are produced in one or a few places. The number of resources is limited. For the region, the quantity of the resources will also affect the price of the resources, and the transportation costs vary from region to region. We need to find the best solution to maximize the profit. This is Resource allocation problem, belonging to nonlinear programing.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 51–62, 2019.
https://doi.org/10.1007/978-3-030-04585-2_7

Resource allocation problem was studied through several methods. David McAllester had shown the nonlinear programing could solve resource allocation problem [1]; Luss et al. used the RAP method to allocate funds for projects that initially consumed cash but later generated cash to maximize the company's return on capital [2]; Karlsson et al. analyzed a cost-benefit of medical resource allocation; Ehrgott et al. proposed a portfolio optimization to create an effective portfolio that allocates funds to stocks or bonds to maximize returns at specific risk levels [4]; Liang et al. used the App_VM_Reconfiguration (AVMR) method to perform dynamic resource configuration based on changes in users' requirements [5].

However, on the scale expanding of production, sales and even resource, the general method is difficult to solve RAP, so many intelligent algorithms were proposed for getting a solution. Hou employed a new genetic algorithm to solve the problem of generalized plant allocation [6]; Kermani [7] proposed a modified version of ant colony optimization for multi-objective resource allocation problem (MORAP), improving the efficiency and effectiveness by increasing the learning of ants in updating pheromone rules and by simplifying probability calculations; Farag [8] proposed a genetic algorithm (GA) based on K-means clustering to solve MORAP.

In this paper a improved differential evolution algorithm was proposed to solve MORAP. The differential evolution algorithm (DE) was proposed by Storn [9] and Price [10]. Its original purpose was to solve the Chebyshev polynomial problem, and later found that the differential evolution strategy can be used in continuous values. Differential evolution differs from genetic algorithms in that offspring are not hybridized by the parent and then mutated. The mutation vector of the differential evolution is generated by the parent difference vector and intersects with the parent individual vector to generate a new individual vector. The unique mutation of DE makes the feature information without problems can also solve the optimization problem in complex environment, being of great significance in transportation, computer science and agriculture etc.

2 Modeling for Allocation Problem

The optimization goal of the allocation problem is to transport the resources to various sales places under the constraint conditions, so that the total profit of these resources in each sales place is the largest, and the objective function is:

$$\max(F) = \max\left(\left(\sum\nolimits_{i=1}^{N} f_i\left(\sum\nolimits_{j=1}^{M} x_{ij}\right)\right) - \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{M} g_{ij}\left(x_{ij}\right)\right) \tag{1}$$

Where N is the number of sale places and M is the number of origin places. The goal is to transport a certain amount of resources from M to N, maximizing the total profit. x_{ij} means resources to the i-th place from the j-th place. The x_{ij} has no compulsion to integers; it could also be a non-integer, for being available to such as beef, soybeans, etc. $f_i\left(\sum_{j=1}^{M} x_{ij}\right)$ is the profit being brought when the total number of resources sold in the i-th place of sale is $\sum_{j=1}^{M} g_{ij}(x_{ij})$. $g_{ij}(x_{ij})$ represents the freight required for the j-th production to deliver the quantity x_{ij} to the i-th sale place.

In recent research, many intelligent algorithms were proposed to solve such resource allocation problem, most of which aim to process discrete problem. Among of the methods, differential evolution algorithm was proposed to solve the allocation problem of continuity.

In the differential evolution algorithm, the solution is often the minimum problem of the objective function. Therefore, when selecting the operation, we need to set $f(u_i(g+1)) \geq f(x_i(g))$, and the profit function and freight function of the product can be obtained from the previous data statistics. In the existing differential evolution algorithm, the value range of the independent variables is relatively independent, and in the transportation allocation problem, there is a constraint relationship between the independent variables. So in the value of the independent variable, we will make the following adjustments.

It is told that the unit shipping cost is commonly different with shipping different quantities of resources; the unit price different for selling different total resources. Therefore, $f_i\left(\sum_{j=1}^{M} x_{ij}\right)$ and $g_{ij}(x_{ij})$ are not linear functions.

Where the number of resource transported from the place of origin j to the sales area i must be greater than or equal to 0, that's

$$x_{ij} \geq 0, \ \ i = 1, \ 2, \ldots, N; \ j = 1, \ 2, \ldots, \ M \tag{2}$$

Since the total amount A_j of a certain product produced in each place is limited, the total output from the place of origin cannot exceed A_j.

$$\sum_{i=1}^{N} x_{ij} \leq A_j, j = 1, 2, \ldots, M \tag{3}$$

When the value of the origin M and the sales ground N reach a tremendous level, the general nonlinear programming becomes difficult to solve; the value of the shipped to the first place in each place of origin is $0 \leq x_{1j} \leq A_j$, and the value ranged to the second place is $0 \leq x_{2j} \leq A_j - x_{ij}$, The value of the value sent to the n-th place is $0 \leq x_{nj} \leq A_j - \sum_{t=1}^{n-1} x_{tj}$.

3 Proposed Method

3.1 Differential Evolution to Solve Allocation Problem

This differential evolution algorithm is divided into four steps. Above all, the population initialization is required, and next the mutation, crossover and selection three-step operations are iteratively executed until the algorithm converges or reaches the number of iterations. First, M individuals are randomly and uniformly generated in the solution space, and each individual is composed of n dimensions.

Initialization. M individuals are randomly and uniformly generated in the solution space, and each individual is composed of n dimensions. As the 0-th generation population, the j-th individual of the i-th individual takes the value as:

$$\begin{gathered} x_{i,j}(0) = L_j + rand(0,1)\left(U_j - L_j\right) \\ i = 1, 2, \cdots, M \\ j = 1, 2, \cdots, n \end{gathered} \tag{4}$$

Where *rand* (0, 1) represents a random number that produces a uniform allocation of *0* to *1*, and U_j and L_j represent the maximum and minimum values of the individual values respectively.

Mutation. The common difference is to randomly select different individuals in the population, and scale the vector difference to perform vector synthesis with the individuals to be mutated. That is:

$$\begin{gathered} V_i(g+1) = x_{r1}(g) + F \cdot (x_{r2}(g) - x_{r3}(g)) \\ i \neq r_1 \neq r_2 \neq r_3 \end{gathered} \tag{5}$$

Where F is the scaling factor, and $x_i(g)$ represents the i-th individual in the g-generation population.

Crossover. The $\{x_i(g)\}$ of the g-th generation population and its variant intermediate $\{v_i(g+1)\}$ are exchanged. The formula is as follows:

$$u_{j,i}(g+1) = \begin{cases} v_{j,i}(g+1), & if\ rand\ (0,1) \leq CR\ or\ j = j_{rand} \\ x_{j,i}, & otherwise \end{cases} \tag{6}$$

Where CR is the crossover probability and j_{rand} is a random integer of [1, 2, …, D].

Selection. The solution of the next generation individual is compared with the solution of the previous generation individual, and the better individual is selected as the individual of the next generation population. That is:

$$x_i(g+1) = \begin{cases} u_i(g+1), & if\ f(u_i(g+1)) \leq f(x_i(g)) \\ x_i(g), & otherwise \end{cases} \tag{7}$$

3.2 Non-linear Random Change Differential Evolution for Allocation Problem

When the DE algorithm is used to solve the allocation problem, the result is likely to fall into the local optimal solution, and the obtained allocation result is commonly not global optimal. Due to the influence of choice, as the evolutionary algebra increases, the greedy algorithm will gradually reduce the differences between individuals. The reduction of individual differences affects the diversity brought about by the mutation, leading to the premature convergence of the algorithm near the extremum. In order to improve the search ability of DE, speed up the convergence and overcome the premature convergence of heuristic algorithms, a non-linear random change differential evolution for multi-objective resource allocation problem was proposed.

Non-linear Control Parameters. Differential evolution algorithm could be improved with the population size NP, the scaling factor F and the crossover probability CR. In the basic differential evolution algorithm, the NP, F, and CR selection methods are usually fixed values from experience, for no changing during the optimization process; furthermore, the performance of the algorithm at various development stages remains challenge as well; therefore, adaptive DE algorithms were generated.

A common adaptive DE algorithm is Chang CS [11] proposed in 2000, being a strategy for linear mutation of scaling factor F and crossover probability CR. Based on this strategy, the scaling factor and crossover probability are improved furtherly. In order to expand the search range at the early stage, the convergence speed is expected to be accelerated. It is known that the crossover probability and scaling factor change slowly in the early iteration, so that one can have a larger search range; in contrast, the medium-term change is faster, so that the individual changes are tend to be slower; therefore, the cross-probability and scaling factor are reduced latterly. The rate of change allows the results to converge to the optimal value. Furtherly, a sigmoid-based function was proposed to adjust the linear mutation of the scaling factor F and the crossover probability CR:

$$CR(g) = \frac{1}{1 - e^{-\left(n - \frac{N}{2}\right)}} * CR_{max} \tag{8}$$

$$F = 1 - F_{max} * \frac{1}{1 - e^{-\left(n - \frac{N}{2}\right)}} \tag{9}$$

Population Structure with Random Change. The population reconstruction is to change the internal parameters of the population under the action of external force when the population has converged. In order to avoid the premature convergence of the population, an adaptive reconstruction method was proposed, called Random Change (RCDE). In the processing of RCDE, the occurrence probability of early random change is large; while of medium-term random change is small.

Two populations are defined at population initialization; one for differential evolution and the other for random numbers within each iteration. The individual number in the two populations are commonly same, and the number with the percentage completed in each iteration plus a constant was compared to a random number. If the number smaller, the individual should be transformed to random individual:

$$f_i = \begin{cases} g_i & if\, \frac{n}{N} + m < rand() \\ f_i & else \end{cases} \tag{10}$$

Where n represents the n-th iteration; N represents the total number of iterations; f_i represents the n-th generation for the i-th individual in the differential evolution population; g_i is the n-th generation for generating the i-th individual in the random number population, and m is a constant that is defined to a range between 0 and 1. In addition, $rand()$ represents a random number between 0 and 1.

In formula (10), the value of m is critical to search the optimal value. If m is too large and close to 1, the occurrence probability of $\frac{n}{N} + m < rand()$ is tend to smallest values, so that the occurrence of $f_i = g_i$ tend to impossibility. This tendency makes the population lose its exploration ability and is prone to premature convergence. In contrast, if the value of m is too small and close to 0, it will make the individual still have a large random mutation in the stage of iteration, so that the population cannot converge.

This method of population reconstruction occurs after the crossover operation and before the selection operation. This mutation could occur in each generation of individuals in the early stage, increasing the diversity of the population. As the number of iterations increases, the probability of such a mutation shows a tendency of decreasing, and after a certain threshold, the individual will not reproduce this mutation, allowing the population to converge to the optimal solution.

4 Experimental Results and Analysis

4.1 Experimental Design

Since the profit function and the price function will be different in different resources, this paper uses a part of the benchmarks listed by Suganthan et al. [12] in 2005. These functions can be approximated as profit objective functions. We tested a total of six functions, 2 of which are unimodal functions, 4 of which are multimodal functions, and the function expressions are shown in Table 1.

Among these six functions, Sphere and Rosenbroke are unimodal functions, and the four functions Ackley, Rastrigin F1, Griewank, and Schaffer F6 are multimodal functions. In this experiment, we conducted three sets of experiments on these six functions. The first set of scaling factor F and the crossover probability CR changed linearly and there was no population reconstruction, and the second group was the scaling factor F and the crossover probability CR changed in S-type and no population remodeling occurs. Finally, the third group has a S-type change of the scaling factor F and the crossover probability CR and a population remodeling occurs, where the value of m is 0.5, that is, after the number of iterations exceeds half of the total number, Population reconstitution no longer occurs.

4.2 Experimental Results

From tests on six different test functions and different dimensions, the experiment results are shown in Table 2. In these tests, the optimal values of the six tested objective functions are all zero, and each generation has 10 individuals; and 50 experiments with 5000 iterations are repeated to obtain the minimum, mean values and

Table 1. Test function expression

Function name	Function expression	Optimal value
Rosenbroke	$f(x) = \sum_{i=1}^{n-1} \left[100\left(x_{i+1} - x_i^2\right)^2 + (x_i - 1)^2\right]$	$\min(f) = f(1, 1, \ldots, 1) = 0$
Sphere	$f(x) = \sum_{i=1}^{n} x_i^2$	$\min(f) = f(0, 0, \ldots, 0) = 0$
Schaffer F6	$f(x) = 0.5 + \frac{\left(sin\sqrt{\sum_{i=1}^{n} x_i^2}\right)^2 - 0.5}{\left(1 + 0.001(\sum_{i=1}^{n} x_i^2)\right)^2}$	$\min(f) = f(0, 0, \ldots, 0) = 0$
Rastrigin F1	$f(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10]$	$\min(f) = f(0, 0, \ldots, 0) = 0$
Ackley	$f(x) = 20 + \mathrm{e} - 20\exp\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \exp(\frac{1}{n}\sum_{i=1}^{n} \cos(2\pi x_i))$	$\min(f) = f(0, 0, \ldots, 0) = 0$
Griewank	$f(x) = 1 + \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\frac{x_i}{\sqrt{i}}$	$\min(f) = f(0, 0, \ldots, 0) = 0$

Table 2. Comparison of experimental results

Function name	Dimension	Linear adaptive mutation			S-type adaptive mutation			Add random change		
		Minimum	Average	Variance	Minimum	Average	Variance	Minimum	Average	Variance
Rosenbroke	10	342767.2	658761.6	1.26E+10	1297.386	34034.71	1.31E+09	**380.4438**	**9053.841**	**70134487**
	20	3090697	4103352	4.25E+11	65984.26	180621.1	9.29E+09	**14663.31**	**88698.1**	**1.64E+09**
	30	7692253	9311698	5.11E+11	243615.4	496289	1.99E+10	**188467.2**	**418558.1**	**1.45E+10**
Sphere	10	4637.485	8010.808	4310562	0.005149	4.1063	32.97037	**7.00E-05**	**1.966284**	**13.55589**
	20	24874.13	30719.29	7837414	3.078489	285.9168	76701.33	**1.596242**	**113.5933**	**9222.241**
	30	59691.26	63797.05	5727514	476.8468	1222.771	312846.6	**421.9374**	**1005.044**	**274888.3**
Schaffer	10	0.491383	0.496576	3.34E-06	0.009716	0.096354	0.00464	**0.009719**	**0.057661**	**0.001871**
	20	0.499485	0.499562	6.91E-10	0.236866	0.413326	0.003777	**0.17828**	**0.33875**	**0.005078**
	30	0.499885	0.499896	1.98E-11	0.45188	0.48541	0.000115	**0.431704**	**0.479907**	**0.000141**
Rastrigin	10	6464.288	8525.17	1777125	3.838911	44.64512	2129.081	**7.546585**	**34.03291**	**520.385**
	20	26954.4	30395.86	3849754	85.95411	338.4269	38796.97	**59.80598**	**233.8338**	**13340.95**
	30	48524.92	54331.69	9485323	476.8468	1222.771	312846.6	**421.9374**	**1005.044**	**274888.3**
Ackley	10	0.379356	0.385354	1.46E-05	0.367879	0.368012	6.87E-08	**0.367879**	**0.36798**	**2.95E-08**
	20	0.4231	0.45275	0.000167	0.367889	0.368514	3.71E-07	**0.367909**	**0.368741**	**6.58E-07**
	30	0.477547	0.518744	0.000217	0.368354	0.370411	1.87E-06	**0.368272**	**0.370929**	**2.67E-06**
Griewank	10	2.993098	3.335894	0.056989	0.025037	0.240268	0.010545	**0.066693**	**0.202522**	**0.006122**
	20	9.161739	9.739362	0.062092	0.194228	0.600499	0.061326	**0.270746**	**0.520103**	**0.03537**
	30	12.98098	15.40648	0.764162	0.99377	1.217582	0.012614	**1.043287**	**1.156992**	**0.009027**

variance. From the results in Table 2, it could be known that the population reorganization is added in the iteration to make the final optimal value much better, no matter whether it's a unimodal function or a multimodal function and whether it's low-dimensional case or multi-dimensional case.

4.3 Search Ability of Parameters for RCDE

In the previous experiments, the value of m was set to 0.5, to keep a constant m in the random change; however, to make a try to get more optimal value of in random change, we furtherly did a series of experiments whose results are shown in Fig. 1. It can be seen from Fig. 1 that the optimal parameters of random change are commonly various in different functions and different dimensions, but it tends to most optimal value when m is around 0.6; and the random change will no longer occur after the number of iterations exceeds 40% of the total number.

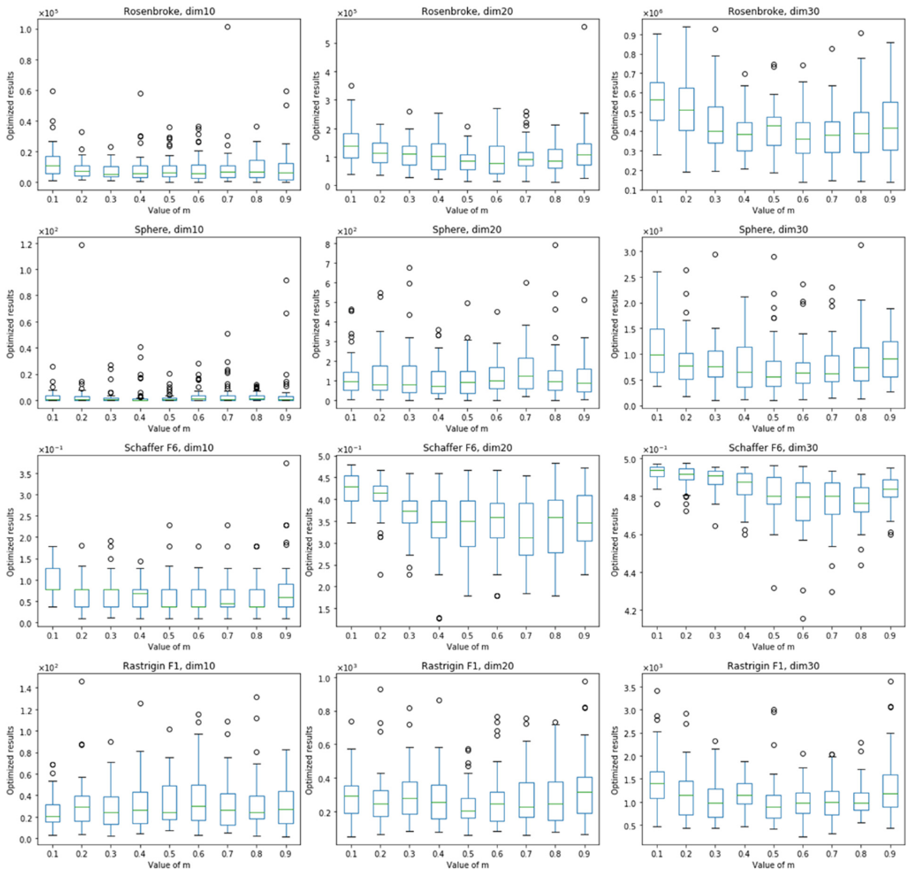


Fig. 1. Box line analysis with different values of m

4.4 Comparison of RCDE Search Capabilities

In order to calculate the exploration ability of random change, we define the distances between individuals, and calculate the distance within the population after each iteration:

$$distance = \frac{1}{2}\left(\sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \left|a_i - a_j\right|\right) \tag{11}$$

Where n is the number of individuals in each iteration, and a is the population individual vector. In a population, the subtraction between two individuals would be double, so the distance between the populations is divided by 2; and the distance of each iteration is shown in Fig. 2.

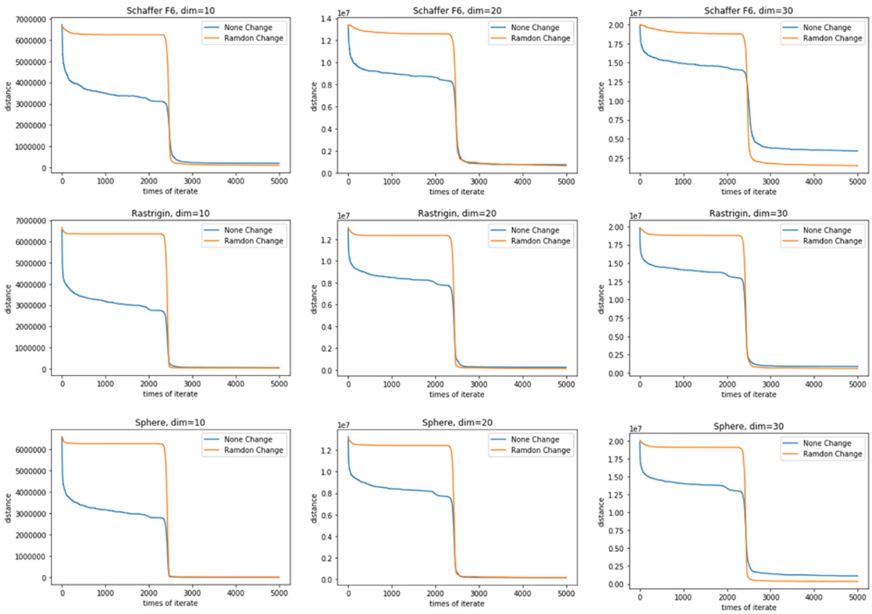


Fig. 2. Inter-individual distance comparison

As could be seen from Fig. 2, the distance of RCDE in the early stage is longer than the individual without RCDE, and in the later stage is less than without RCDE. The experiments results show that the individuals with RCDE have more ability to explore in the early stage, and also have more ability to converge to the optimal value in the later stage. Especially, when dimension increasing, this advantage is more obvious.

5 Conclusion

To deal with problem of multi-objective resource allocation, a novel differential evolution algorithm was proposed to find the optimal solution. For improving the exploration ability of differential evolution, the novel method is of adaptability in scaling factor and crossover probability, making the population closer to the optimal solution. At the same time, an algorithm transforming the individual in each generation into another random one in each iteration was proposed to reduce the premature convergence of the population. Finally, the method was evaluated using six functions in 10, 20 and 30 dimensions, and the results including mean, standard deviation and optimal value show that the improved DE algorithm is superior to other existing methods in optimization accuracy.

Acknowledgment. The authors thank Prof. Jianhua Liu with the Fujian Provincial Key Lab of Big Data Mining and Applications at the Fujian University of Technology for valuable discussions on this study.

This work was supported in part by Projects of National Science Foundation of China (No. 41471333); project 2017A13025 of Science and Technology Development Center, Ministry of Education, China; project 2018Y3001 of Fujian Provincial Department of Science and Technology; projects of Fujian Provincial Department of Education (JA14209, JA15325).

References

1. McAllister, D., Rosenblitt, D.: Systematic nonlinear planning. In: Proceedings of AAAI-91, Anaheim (1991)
2. Luss, H., Gupta, S.K.: Allocation of effort resource among competing activities. Oper. Res. **23**, 360–366 (1975)
3. Karlsson, G., Johannesson, M.: The decision rules of cost-effectiveness analysis. PharmacoEconomics **9**, 113–120 (1996)
4. Ehrgott, M., Klamroth, K., Schwehm, C.: An MCDM approach to portfolio optimization. Eur. J. Oper. Res. **155**(3), 752–770 (2004)
5. Liang, Q., et al.: The placement method of resources and applications based on request prediction in cloud data center. Inf. Sci. **279**, 735–745 (2014)
6. Hou, Y.C., Chang, Y.H.: A new efficient encoding mode of genetic algorithms for the generalized plant allocation problem. J. Inf. Sci. Eng. **20**, 1019–1034 (2004)
7. Chaharsooghi, S.K., Kermani, A.H.M.: An effective ant colony optimization algorithm (ACO) for multi-objective resource allocation problem (MORAP). Appl. Math. Comput. **200** (1), 167–177 (2008)
8. Mousa, A.A., El-Shorbagy, M.A., Farag, M.A.: K-means-clustering based evolutionary algorithm for multi-objective resource allocation problems. Appl. Math. Inf. Sci. **11**(6), 1–12 (2017)
9. Storn, R., Price, K.: Differential Evolution - A Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces. University of California, Berkeley (2006)

10. Price, K.V., Stone, R.M., Lampinen, J.A.: Differential Evolution: A Practical Approach to Global Optimization. Springer, Heidelberg (2005)
11. Chang, C.S., Xu, D.Y.: Differential evolution based tuning of fuzzy automatic train operation for mass rapid transit system. IEEE Proc-Electric Power Appl. **147**(3), 206–212 (2000)
12. Suganthan, P.N., Hansen, N., Liang, J.J., Deb, K., Chen, Y.P., Auger, A., et al.: Problem definitions and evaluation criteria for the CEC2005 special session on real-parameter optimization. Technical report, Report #2005005. Nanyang Technological University, Singapore (2005)

Trajectory Similarity Measuring with Grid-Based DTW

Qiqin Cai[1,2], Lyuchao Liao[1,2](✉), Fumin Zou[1,2], Subin Song[3], Jierui Liu[1,2], and Meirun Zhang[1]

[1] Fujian Key Lab for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China
cail994022801l1@gmail.com, fjachao@gmail.com, fmzou@fjut.edu.cn
[2] Fujian Provincial Big Data Research Institute of Intelligent Transportation, Fujian University of Technology, Fuzhou 350118, Fujian, China
[3] Fujian Transport Information and Telecommunication Center, Fujian Communication Department, Fuzhou 350001, Fujian, China

Abstract. With the rapid accumulation of GPS trajectory data, the vast amount of spatiotemporal trajectory data hides extremely rich and valuable information with potential travel behavior patterns. The similarity of the driver's travel trajectory is key to mining patterns, but how to reasonably and efficiently evaluate the similarity remains a challenge. To address this problem, we propose a driving trajectory similarity measurement using grid-based dynamic time warping (GDTW) to evaluate similarity of driving trajectory. Building trajectory grid vector model (TGVM), the method solves the problems of position shift and large computation for the similarity measuring of big trajectory data. Extensive experiments were conducted with a real trajectory dataset to evaluate feasibility and efficiency of the proposed approaches. The results show that GDTW performs a more robust and efficient processing of trajectory similarity than does traditional approaches, reducing by about 5 times of time-consuming.

Keywords: Trajectory similarity measuring · DTW · Trajectory data mining Driving behavior cluster

1 Introduction

In recent years, information technologies have significantly penetrated transportation systems. Especially, increasingly vehicles are being equipped with positioning devices, which continuously generate spatiotemporal trajectory data; therefore, numerous driving trajectories are accumulated. These massive spatiotemporal trajectories could be studied to understand driving behavior patterns [1, 2]. Using these driving behavior patterns, traffic information can be provided more efficiently, and the whole traffic flow could also be predicted. For example, a traffic-flow density cluster identifying similar trajectories shows indicators of emerging traffic measures, future road expansions, traffic jam detection, traffic predictions, and so on [3]. Further, trajectory similarity can also be used in numerous road network applications such as city emergency

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 63–72, 2019.
https://doi.org/10.1007/978-3-030-04585-2_8

management, driver guidance systems, and path-recommendation systems supporting analysis of individual data [4]. Therefore, an efficient trajectory similarity measurement, being the primary step in mining underlying patterns from massive trajectory data, is of much requirement [5].

Current related studies could be summarized as Euclidean, dynamic time warping (DTW), longest common subsequence (LCSS), edit distance on real sequences (EDR), and discrete Fourier transform (DFT). Most of these methods are aim to make a precise matching, relying on point matching. However, trajectory data indicating driving behavior involve many actual affect factors including time asynchronous, position shift errors and transmission losing; the data are also typically of sparsity; these factors make it hard to get a precise matching in similarity measuring [6]. Furtherly, these current methods are lack of efficiency to process big trajectory data. Therefore, a more efficient and feasible method is required urgently to process massive trajectory data.

To address these problems, basing on trajectory grid vector model (TGVM) [1], a novel grid-based DTW algorithm (GDTW) was proposed to evaluate the similarity between trajectories, combining TGVM and traditional DTW algorithm to reduce data dimension and correct position shift error. This method was tested with a real dataset and showed efficiency and effectivity of significance.

The rest of this paper is organized as follows. First a brief review of existing related works is presented in Sect. 2. Section 3 presents the modeling method based on trajectory grid vector model (TGVM), and proposes a novel GDTW, combining TGVM to traditional DTW, to distinguish similar trajectories from massive spatiotemporal trajectory data. Experimental tests with real trajectory dataset for demonstrating performance and effectivity are presented and analyzed in Sect. 4. Finally, we conclude our work and look forward to future work in Sect. 5.

2 Related Works

Several related studies have investigated the different methods of measuring the similarity of spatiotemporal trajectories. The primitive similarity metric is Euclidian distance measurement [7, 8], which is of less complex than the other methods; however, it works only on trajectories of same duration and granularity. Moreover, Euclidean-based methods are hard to handle sparse and asynchronous trajectory data, especially those with noise and gaps.

Therefore, several more robust similarity-measuring methods, such as EDR [9, 10], LCSS [11, 12], DTW [13, 14], DFT [15] and so on, have been proposed. The LCSS is a variation of the edit distance method and is defined as the number of consecutive common locations between two trajectories. The basic idea of LCSS is to match two sequences by allowing them to stretch, without rearranging the sequence of elements, and allowing some elements to be unmatched. The LCSS is advantageous because it can address issues related to trajectories taken at different spatial granularities and is efficient even if the trajectory contains a significant amount of noise data. However, this method does not take temporal or spatiotemporal variables into consideration.

The edit distance method is based on a new representation of trajectories, referred to as movement pattern strings (MPS). In the MPS scheme, a trajectory is represented using movement direction and distance ratio information derived from original trajectory. This method is more accurate than the other commonly used trajectory similarity measures, particularly in the presence of noise in data. To optimize the similarity computation, several extensions of the edit distance method have been developed. For example, a distance function based on edit distance on real sequence (EDR) was introduced by Chen et al. [16].

These existing methods could effectively measure similarity between two trajectories; however, they are lack of processing efficiency and hard of measuring similarity from massive trajectory data with high dimensions.

3 Proposed Method

A spatial trajectory, concatenating location points, is a sequence of coordinate value pairs tagged with timestamps. For efficient similarity measuring, the following challenges must be considered when modeling trajectory data. (1) A trajectory is commonly of sparsity due to aiming to reduce transmission cost; (2) the same driving route could show obvious differential trajectories sequence, being of much asynchronism, position shift, and loss of packet transmission; (3) non-equal length is a natural character of trajectory data, expecting to be overcome to improve processing efficiency, especially for massive data.

The method to measure trajectory similarity should be able to overcome the challenges above, so the TGVM modeling method [1] was introduced to combine traditional DTW method [17]. TGVM is based on the spatial grid model to reduce data accuracy and improve processing efficiency, considering each trajectory is identified by a sequence of passing grids. Based on TGVM, a novel grid-based dynamic time warping algorithm (GDTW) was proposed to efficiently measure the trajectory similarity from massive trajectory data.

Definition 1 Spatiotemporal Trajectory. Vehicles equipped positioning devices periodically report their location information $p(lon, lat, t)$, and generate trajectories using time-indexed sequences of these multidimensional information. The trajectory representation is shown as follow:

$$TR_i = <p_1, p_2, \ldots, p_j, \ldots, p_{ni}> \tag{1}$$

Where $p_j[1 \leq \mathrm{j} \leq len(TR_i)]$ is a multi-dimensional location, and the length n_i of a trajectory is commonly different from each other.

3.1 Trajectory Grid Vector Modeling

The core idea of trajectory grid vector modeling is to reduce data accuracy and data dimension. Position apparatus commonly generate longitude/latitude at the utmost accuracy, making data, even at the same position, different from each other. We introduce the trajectory grid vector model (TGVM) to reduce uncertainty of data and meet the massive trajectory mining requirement.

With complete road network space partitioned into meshing grids, any location that is a part of the trajectory can be mapped into one and only one grid; therefore, the trajectory can be represented with a sequence of passed spatial grids.

$$TR_i = < g_{i1}, g_{i2}, \ldots, g_{ij}, \ldots, g_{in} > \quad (2)$$

By tuning size of grids, the accuracy of trajectory can be adjusted. As shown in Fig. 1, the trajectory TR_1, containing 20 points, is represented by nine grids (TR_1 = <g1, g6, g7, g8, g9, g14, g13, g18, g17, g16>).

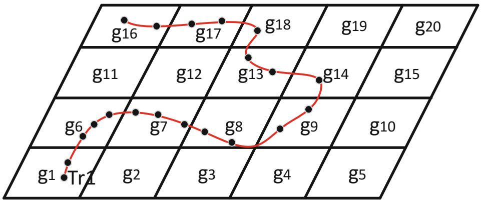


Fig. 1. Trajectory representation with spatial grid

Thus far, the infinite coordinates have been mapped onto a small number of grids, which also makes the position shift endurable.

3.2 Grid-Based Dynamic Time Warping

Aligning two time series by extending or contracting, dynamic time warping (DTW) is feasible of evaluating the similarity of time sequences with different length. Supposed query trajectory: $TR^{query} = < q_1, q_2, \ldots, q_m >$, m is number of trajectory points in query trajectory; target trajectory:$TR^{target} = < p_1, p_2, \ldots, p_n >$, n is number of trajectory points in target trajectory. The result that point pair matching between TR^{query} and TR^{target} is as follows (Fig. 2).

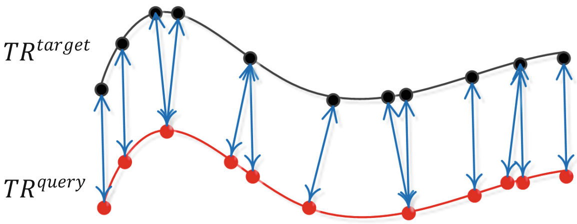


Fig. 2. Point pair matching between TR^{query} and TR^{target}

For getting match with simplicity and efficiency, traditional Euclidean distance was employed to calculate similarity between TR^{query} and TR^{target}.

$$DTW\left(TR^{query}, TR^{target}\right) = D(m, n) \tag{3}$$

Where,

$$D(\mathrm{i},\mathrm{j}) = Dist\left(q_i - p_j\right) + min\{D(i-1,j), D(i,j-1), D(i-1,j-1)\} \tag{4}$$

$$Dist\left(q_i - p_j\right) = \sqrt{\left(lon_i - lon_j\right)^2 + \left(lat_i - lat_j\right)^2} \tag{5}$$

Where $Dist()$ denotes the Euclidean Distance of two point; i, j denotes i-th and j-th point in query trajectory and target trajectory, respectively.

To evaluate the best matching effect, dynamic programming (DP) is used to warp path, so that the distance of all matching points is shortest (Fig. 3).

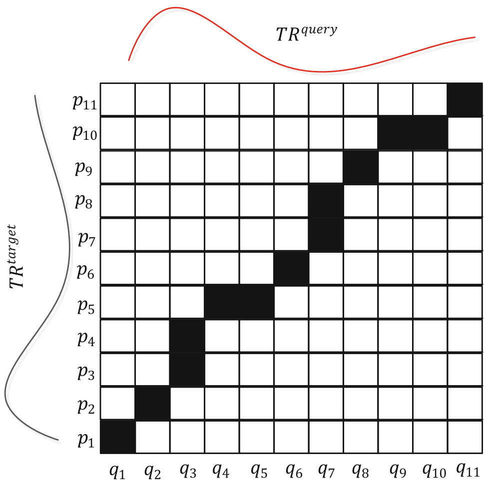


Fig. 3. Warping path

Dynamic time warping (DTW) make tries to get warping solution for different sampling period and asynchronism, while failing to deal with massive dataset. With the trajectory grid vector model [1] employed, a novel grid-based dynamic time warping (GDTW) to evaluate the similarity between trajectories. GDTW inherit the advantages both of TGVM to correct position shift error and reduce data dimension and of DTW to make warping for asynchronism. As shown in Fig. 4, GDTW matching between query trajectory and target trajectory is transformed to grid pair matching without synchronism, on the query trajectory not in step with the target trajectories.

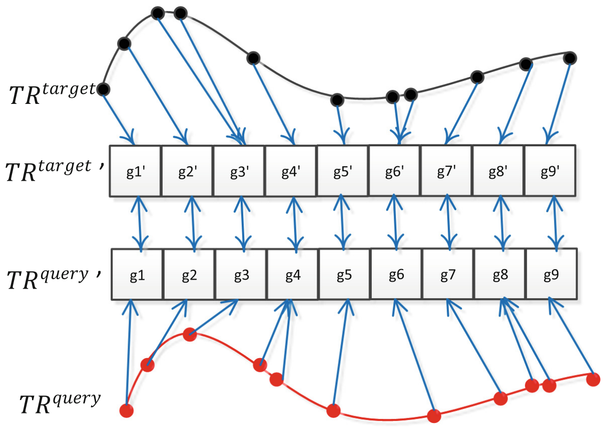


Fig. 4. Grid pair matching between TR^{query} and TR^{target}

4 Experimental Results and Analysis

The proposed method was evaluated with a real trajectory dataset obtained from the Fujian Communications Department of China, which has accumulated a large amount of trajectory data. The space area is limited to longitude/latitude ranges of ([83, 123], [18, 53]). The experiments were conducted on a DELL PC equipped with Intel Core running at 2.6 GHz and 128 GB RAM and programmed in R.

4.1 Effectiveness Tests

For this experiment, a user trajectory was randomly selected from the real trajectory dataset, which has three months of trajectory data and can be divided into 611 sub-trajectories. When query trajectory TR_1^{query} randomly selected from 611 sub-trajectories is selected, the remaining 610 sub-trajectories will be used as the target trajectory TR^{target}, so and on, for each sub-trajectory as query trajectory. Subsequently, similar trajectories from the whole dataset was conducted by the similarity evaluation of each other.

To make illustration of experimental results, two trajectories shown in Fig. 5(a) is randomly selected as query trajectory; the Fig. 5(b) and (c) show their top 5 most similar trajectories in the experimental dataset, in which the overlapping trajectories retrieved were shifted around to make a better visible illustration.

As shown in Fig. 5, the shape of 5 trajectories retrieved are evidently coincident and consistent with query trajectory TR^{query}. And from detail drawing in Fig. 5(b) and (c), it can be seen that the 5 retrieved trajectory data are not coincident with accurate point location but with position drifting, while the proposed algorithm GDTW as well as DTW could distinguish their similarity with effectivity.

The Euclidean distance between query trajectory and respective retrieved trajectories was computed to evaluate the effectivity of GDTW. As shown in the Fig. 6, the experiments made computation of Euclidean distance to the top 50 most similar trajectory retrieved by GDTW and DTW.

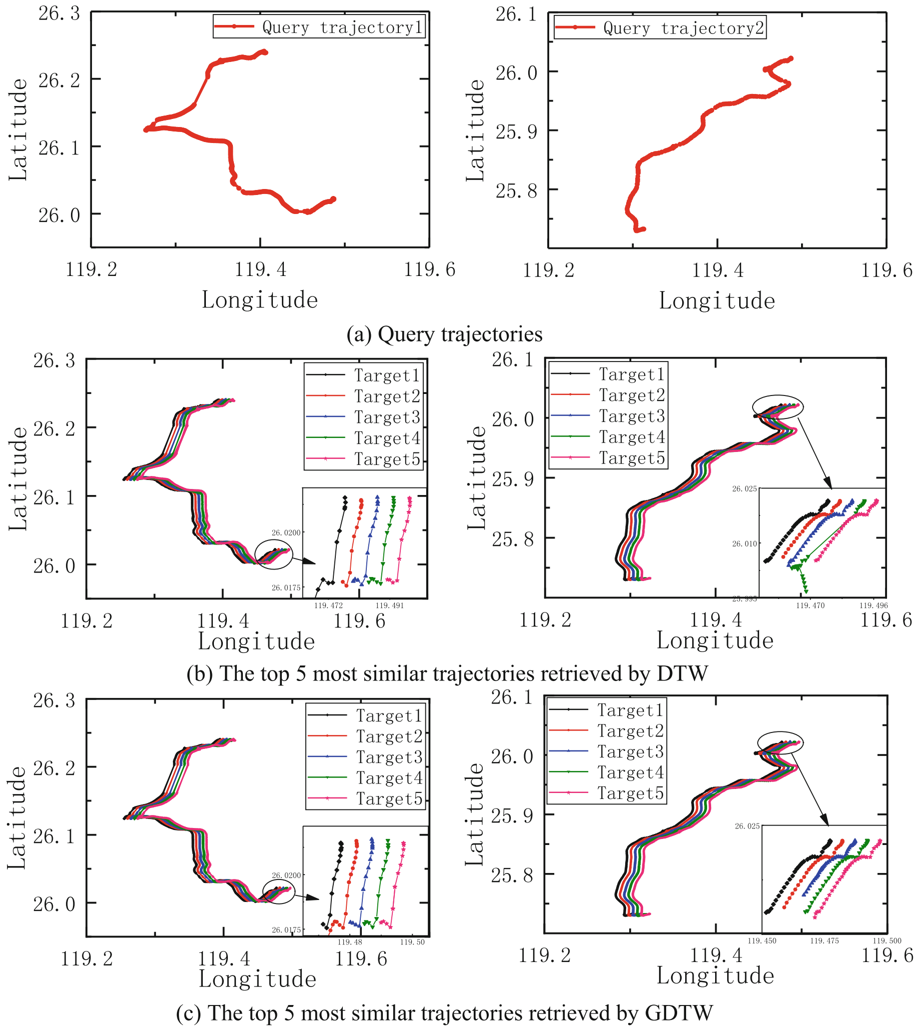
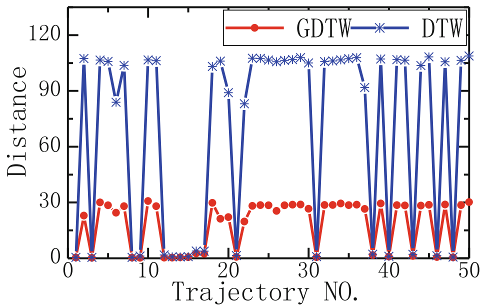


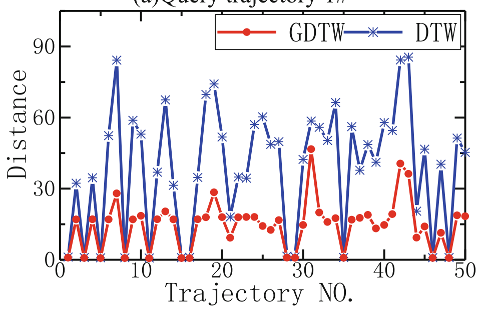
Fig. 5. Query trajectories and retrieved results of DTW and GDTW

From the comparison of Euclidean distance, the distance of results with GDTW are commonly lower than does DTW being of no dimension reduction. That is,

$$DTW(TR^{query}, TR^{target}) \geq GDTW(TR^{query}, TR^{target})$$



(a)Query trajectory 1#



(b) Query trajectory 2#

Fig. 6. Euclidean distance of top 50 most similar trajectories with DTW and GDTW

The lower Euclidean distance in GDTW means having more ability to detect similar trajectories, especially in massive trajectory dataset, and means having better adaptation to various trajectory while filtering by threshold set manually.

4.2 Efficiency Tests

Time consuming was employed to evaluate the efficiency of GDTW. Firstly, five trajectories randomly selected from 611 sub-trajectories were employed as target trajectories TR^{target}, while keeping TR^{query} fixed. In these experiments, the TR_1^{query} contains 1609 points at total and TR_2^{query} 575 points; the data counting of target trajectories are listed in Table 1.

As can be seen from Table 1, with the increasing of data number, the time consuming of both DTW and GDTW are increasing, but comparing with DTW, the proposed GDTW is more efficient, reducing average time consuming by about 5 times. Meanwhile, the time-consuming ratios between GDTW and DTW, showing a tendency of more efficiency in bigger dataset, also illustrates the efficient processing power of GDTW for large-scale data.

Table 1. Time-consuming comparison between DTW and GDTW

(a) Query trajectory 1#				
TR^{target}	Data counting	Time consuming/s		Time-consuming ratio (DTW/GDTW)
		DTW	GDTW	
1	1093	36.74	6.72	5.47
2	1333	45.12	8.37	5.39
3	1514	53.37	7.26	7.35
4	1918	70.08	11.01	6.36
5	2591	86.88	12.74	6.82
(b) Query trajectory 2#				
TR^{target}	Data counting	Time consuming/s		Time-consuming ratio (DTW/GDTW)
		DTW	GDTW	
1	1093	13.68	3.17	4.32
2	1333	17.06	4.39	3.74
3	1514	18.91	4.58	4.12
4	1918	25.48	5.87	4.34
5	2591	32.46	6.66	4.87

5 Conclusion

An efficient similarity measurement is one of the most primary steps in trajectory data mining, especially for underlying driving behavior patterns from massive data. However, the existing methods are time-consuming and unsuitable for massive trajectory dataset. To address this challenge, a grid-based dynamic time warping (GDTW) is proposed to evaluate driving-trajectory similarity, employing a trajectory grid vector model to reduce data dimensions and time complexity. Being non-sensitive to problems such as position drift and asynchronous, GDTW makes it possible to quickly discover similar trajectories with efficiency. Experimental results show that GDTW is of efficiency and robust to process increasing trajectory dataset.

To be more adaptability for various dataset, there still exist some challenges in GDTW. The grid size determines directly the accuracy and efficiency of the method. Being of smaller size, the accuracy increases but the efficiency decreases, and vice versa.

Acknowledgment. This work was supported in part by Projects of National Science Foundation of China (No. 41471333); project 2017A13025 of Science and Technology Development Center, Ministry of Education, China; project 2018Y3001 of Fujian Provincial Department of Science and Technology; projects of Fujian Provincial Department of Education (JA14209, JA15325). Fujian Provincial Department of Transportation is also acknowledged for supporting the experimental dataset.

References

1. Liao, L., Wu, J., Zou, F., et al.: Trajectory topic modelling to characterize driving behaviors with GPS-based trajectory data. J. Internet Technol. **19**(3), 815–824 (2018)
2. Carboni, EM., Bogorny, V.: Inferring drivers behavior through trajectory analysis. In: Intelligent Systems 2014, pp. 837–848. Springer, Cham (2015)
3. Shang, S., Chen, L., Wei, Z., et al.: Trajectory similarity join in spatial networks. Proc. VLDB Endow. **10**(11), 1178–1189 (2017)
4. Li, Y., Su, H., Demiryurek, U., et al.: PaRE: a system for personalized route guidance. In: Proceedings of the 26th International Conference on World Wide Web, pp. 637–646. International World Wide Web Conferences Steering Committee (2017)
5. Long, J.A., Nelson, T.A.: A review of quantitative methods for movement data. Int. J. Geograph. Inf. Sci. **27**(2), 292–318 (2013)
6. Sankararaman, S., Agarwal, P.K., Mølhave, T., et al.: Computing similarity between a pair of trajectories. arXiv preprint arXiv:1303.1585 (2013)
7. Atev, S., Miller, G., Papanikolopoulos, N.P.: Clustering of vehicle trajectories. IEEE Trans. Intell. Transp. Syst. **11**(3), 647–657 (2010)
8. Kahveci, T., Singh, A., Gurel, A.: Similarity searching for multi-attribute sequences. In: Proceedings of the 14th International Conference on Scientific and Statistical Database Management, pp. 175–184. IEEE (2002)
9. Furtado, A.S., Kopanaki, D., Alvares, L.O., et al.: Multidimensional similarity measuring for semantic trajectories. Trans. GIS **20**(2), 280–298 (2016)
10. Toohey, K., Duckham, M.: Trajectory similarity measures. Sigspatial Spec. **7**(1), 43–50 (2015)
11. Kang, H.Y., Kim, J.S., Li, K.J.: Similarity measures for trajectory of moving objects in cellular space. In: Proceedings of the 2009 ACM Symposium on Applied Computing, pp. 1325–1330. ACM (2009)
12. Lin, B., Su, J.: One way distance: For shape based similarity search of moving object trajectories. GeoInformatica **12**(2), 117–142 (2008)
13. Lim, E.C., Shim, C.B.: Similarity search algorithm for efficient sub-trajectory matching in moving databases. In: International Conference on Computational Science, pp. 821–828. Springer, Heidelberg (2007)
14. Sakurai, Y., Yoshikawa, M., Faloutsos, C.: FTW: fast similarity search under the time warping distance. In: Proceedings of the Twenty-Fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 326–337. ACM (2005)
15. Hu, W., Li, X., Tian, G., et al.: An incremental DPMM-based method for trajectory clustering, modeling, and retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **35**(5), 1051–1065 (2013)
16. Chen, L., Özsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 491–502. ACM (2005)
17. Itakura, F.: Minimum prediction residual principle applied to speech recognition. IEEE Trans. Acoust. Speech Signal Process. **23**(1), 67–72 (1975)

Dynamic Response of Levitation Force to Magnetic Field Fluctuation for High Temperature Superconducting Maglev System

Lifeng Zhao[1,2], Jiangtao Deng[1,2], Linbo Li[1,2], Jing Jiang[1,2], Yong Zhang[1,2], and Yong Zhao[1,2,3](✉)

[1] Key Laboratory of Magnetic Levitation Technologies and Maglev Trains, Ministry of Education of China, Southwest Jiaotong University, Chengdu 610031, China
{zhaolf,yzhao}@home.swjtu.edu.cn
[2] Superconductivity and New Energy R&D Center, Southwest Jiaotong University, Chengdu 610031, China
[3] College of Physics and Energy, Fujian Normal University, Fuzhou 350117, Fujian, China

Abstract. Dynamic response of high temperature superconductor (HTS) to inhomogeneous distribution of magnetic field for permanent magnetic guideway (PMG), as well as to perturbations from the magnetic field of the PMG partially covered by iron sheet, are investigated. Decay of levitation force is observed caused by AC magnetic field. Furthermore, resonant effect at 180 Hz is obtained caused by inhomogeneous distribution of magnetic field of permanent magnet guideway. Otherwise, electromagnetic oscillations caused by abruptly decrease of magnetic field are also observed. These are perhaps related to complicated electromagnetic interaction inside HTS bulk.

Keywords: Dynamic response · High temperature superconductor Maglev

1 Introduction

Since the discovery of HTS, it has attracted extensive attentions for its potential application on maglev train [1, 2]. The unique Meisner effect and flux-pinning properties of HTS make it levitate over permanent magnets without extra control system [3]. Much effort has been done to promote the practical application [4–6], and numerous of remarkable achievements have also been realized. However, it still needs a long time for the real commercial application because many key technique problems still need to be solved. Among them, the decay of levitation performance caused by AC external magnetic field is the most important.

The inhomogeneous magnetic field distribution and installation error of the PMG, or flexure deformation of viaduct with guideway lying on will cause high or low frequency AC magnetic field for HTS bulks to experience when maglev train is running at a high speed along PMG. Much effort has been done to discuss the dynamic performance of maglev system in traditional ways [7–9], i.e., mechanical vibration or

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 73–79, 2019.
https://doi.org/10.1007/978-3-030-04585-2_9

alternating magnetic field from a solenoid magnet was taken as the external excitation source to study the dynamic performance of maglev system. It is different with the situation of inhomogeneous distribution of magnetic field experience by maglev vehicle. For the reason, with the increase of the speed, the changing rate of the magnetic field (*dB/dt*) experienced by the maglev system can increase in a larger scale, which is difficult for traditional mechanical vibration or solenoid magnet to reach.

Different with some researchers' discussion on the dynamic response for HTS maglev moving along PMG [10, 11], our early results reported the effects of amplitude of alternating magnetic field and *dB/dt* on the dynamic response of levitation performance of HTS maglev system [12], but is still insufficient. In this letter, we discuss the effects of different amplitude of alternating magnetic field on the dynamic response of levitation performance. Some phenomena observed in our experiment are perhaps related to electromagnetic interaction inside HTS bulk. It should be helpful for the study of HTS in the future.

2 Measurement and Experiment

The experimental device is shown in Fig. 1, which had been reported in our earlier report [12]. The circular PMG with the diameter of 1 m is consisted of 24 circular permanent magnets, and it is driven by a motor to rotate horizontally. In order to change the magnetic field distribution artificially, two iron sheets in thickness of 1, 2 and 3 mm are symmetrically arranged on the circular guideway, respectively. The center angle for each iron sheet is 100°. Figure 2 shows the profile of the measurement result of the vertical component of the magnetic field of the PMG circumferential field at the height of 5 mm. As a magnetic material, iron sheet covering on the PMG can effectively reduce intensity of magnetic field above its surface.

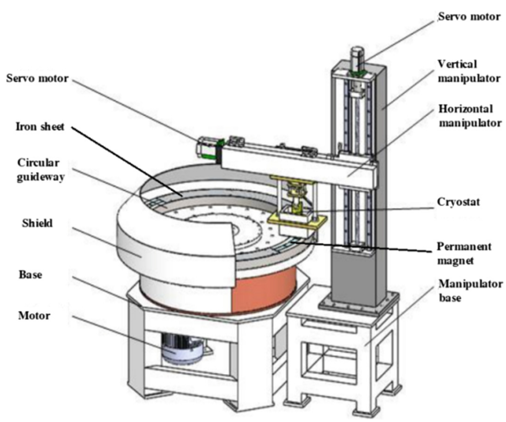


Fig. 1. The measurement scene of the HTS Maglev dynamic test system

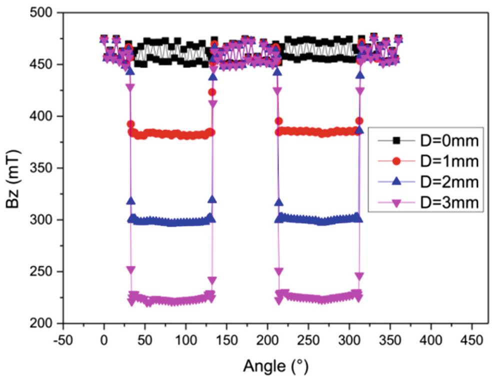


Fig. 2. Magnetic field distribution of PMG with iron sheets in different thickness at the height of 5 mm.

A melt-textured YBCO bulk with the diameter of 30 mm and thickness of 18 mm is used in the experiment. The HST bulk was field cooled at the gap of 30 mm above the PMG region, and then pressed to the working height of 5 mm to hold for 10 min. Then the PMG with iron sheets (in different thickness of D = 0, 1, 2, 3 mm) was accelerated to rotate successively from zero to 110 km/h in 140 s, then kept at the speed of 110 km/h for 70 s, and finally decelerated back to zero in about 140 s, respectively.

3 Results and Discussion

The typical dynamic responses of levitation performance for the PMG with iron sheet in thickness of 0 and 3 mm are shown in Fig, 3, respectively. We can observe that, at the accelerating stage for the situation of D = 0, the amplitude of levitation force increases gradually and reaches the maximum at the speed about 80 km/h, then decreases abruptly. At the constant speed stage of 100 km/h, the amplitude of levitation is almost a constant. While in the decelerating stage, the variation of amplitude of is almost symmetrical with that of the accelerating stage.

The maximum observed near 80 km/h seems to be related to resonant behavior. As our circular PMG is consisted of 24 permanent magnets, the intensity of magnetic field over the joint between two adjacent magnets is smaller than that of PMG elsewhere. Furthermore, there exists inhomogeneous distribution of magnetic field for every magnet caused by the limited resolution during its fabrication and magnetization factor. Thus the rotated PMG will cause the bulk to experience AC magnetic field. Considering the alternating magnetic field caused by joints, the frequency of AC magnetic field experienced by bulk is 180 Hz at the speed of 80 km/h. However, earlier results discussed that resonant effect is often observed at a frequency below 20 Hz for a free suspended HTS maglev [13]. Some reporters further pointed out that the resonant frequency is depending on suspending height (levitation stiffness) and suspended mass [14]. In our experiment, the bulk did not suspend but be fixed over PMG. Thus the resonant effect observed perhaps is related to complicated electromagnetic interaction inside HTS bulk. Nevertheless, similar phenomenon was not observed in our earlier experiments with the working height of 10 mm [12]. We suggest that the intensity of magnetic field and the amplitude of its AC component decrease with the increase of working height because of the mutual exclusion among fluxes. Thus it will not cause the resonant effect similar to that of the working height of 5 mm.

For the case of D = 3 shown in Fig. 3, the amplitude of AC magnetic field is 250 mT, which is much larger than that of D = 0. Thus it causes larger amplitude of levitation force. As two iron sheets are symmetrically arranged on the surface of PMG, its related frequency is 15 Hz at the speed of 80 km/h. Furthermore, we can observe that the amplitude of levitation force augments with the increase of speed, and diminishes with the decrease of speed. However, no apparent resonant effect can be figured out. The frequency of 15 Hz is much less than that of D = 0 of 180 Hz, while the amplitude of levitation caused by iron sheet with D = 3 is much larger than that caused by inhomogeneous distribution of magnetic field among naked sections of PMG. Thus the resonant effect similar to that of D = 0 cannot be observed apparently.

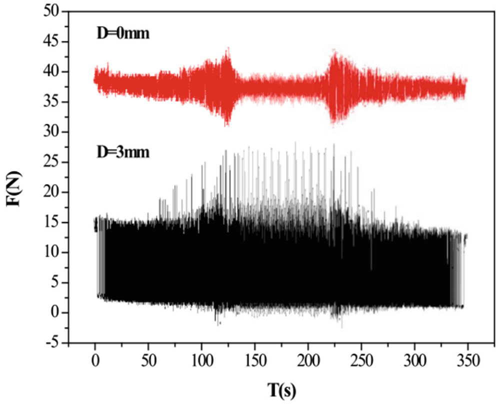


Fig. 3. Variation of levitation force with speed for PMG covered with iron sheet in thickness of 0 and 3 mm.

Figure 4 presents the variation of levitation force for PMG covered with iron sheet in thickness of 3 mm at different speed. It is similar to the results presented before [12]. With the increase of speed, changing rate of the magnetic field (*dB/dt*) experienced by the maglev system increases at the same time. This causes an increasing sharp peak in the profile of levitation force, especially for the case of $dB/dt > 0$ (where the bulk moves from iron sheet to the PMG). As the iron sheet reduces the inhomogeneity of magnetic field distribution, as shown in Fig. 2, the fluctuation of levitation force for the region covered with iron sheet is less than that for the region without iron sheet (D = 0). Furthermore, for the region covered with iron sheet, we can observe the amplitude of levitation force decrease gradually after the bulk's crossing the junction from PMG to iron sheet. Qiu reported similar phenomenon and contribute it to electromagnetic oscillations caused by instantaneous impact [15].

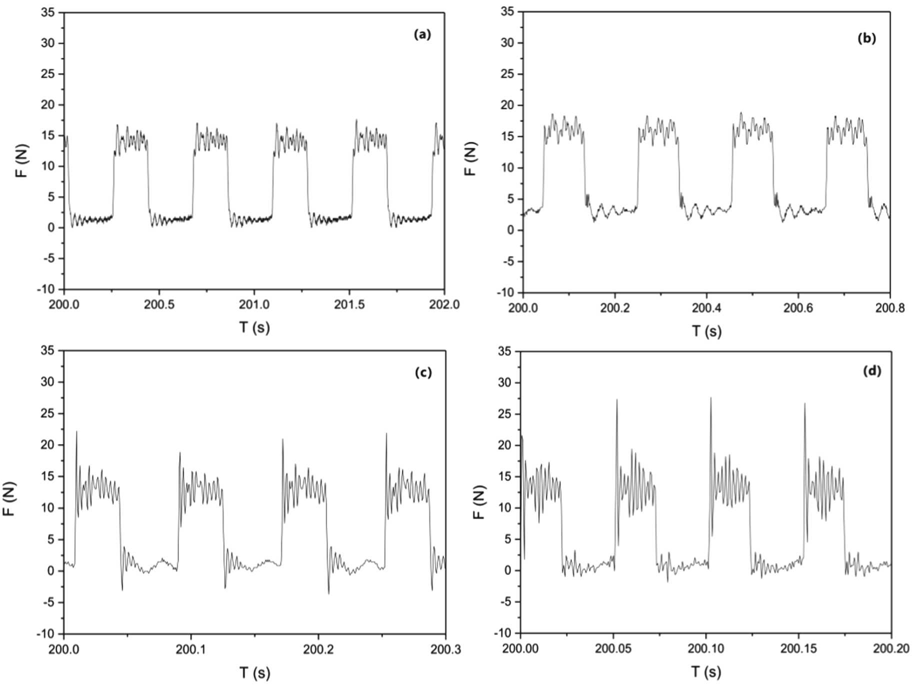


Fig. 4. Variation of levitation force for PMG covered with iron sheet in thickness of 3 mm at different speed. (a) 10 km/h. (b) 27 km/h. (c) 68 km/h. (d) 110 km/h.

The attenuations of levitation force for the situations of PMG covered with iron sheets in thickness of 0, 1, 2 and 3 mm are obtained by comparing the levitation force before (F_0) and after (F_1) the rotation of PMG. The results are presented in Fig. 5. As the inhomogeneous distribution of magnetic field for PMG, as well as perturbations from the magnetic field of the PMG partially covered by iron sheet, can be considered as AC magnetic field to be experienced by bulk. This will cause AC losses in bulk and further the attenuation of levitation force. The larger the amplitude of AC magnetic field is, the greater the attenuation of levitation force will be presented [16]. Thus the increased amplitude of magnetic field caused by the increase of thickness of iron sheet augments the decay of levitation force.

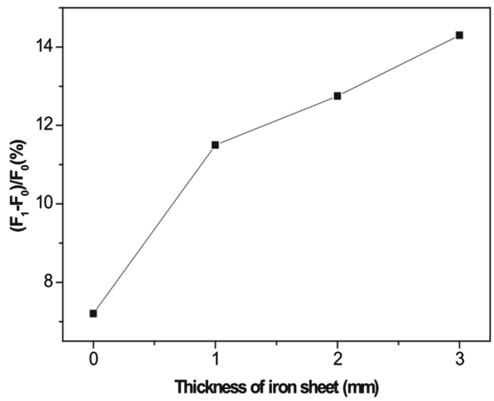


Fig. 5. Decay rate of levitation force for iron sheet with different thickness of 0, 1, 2 and 3 mm.

4 Conclusion

Dynamic response of HTS bulk to inhomogeneous distribution of magnetic field for PMG, as well as perturbations from the magnetic field of the PMG partially covered by iron sheet, are investigated. Decay of levitation force caused by AC magnetic field is observed. Otherwise, for the situation of naked PMG, resonant behavior can be observed at about 180 Hz. Furthermore, for the situation of PMG partially covered with iron sheet, electromagnetic oscillations are observed after the bulk's crossing the junction from PMG to iron sheet. Such phenomena perhaps imply complicated electromagnetic interaction inside HTS bulk. It should be helpful for the study on HTS in the future.

Acknowledgements. This work was supported in part by the fund for the project support, including the Sichuan Applied Basic Research Project (No. 2018JY0003, 2017JY0057), the Program of International S&T Cooperation (Grant No. 2013DFA51050), the Fundamental Research Funds for the Central Universities under No. 2682016 ZDPY10 and Foundation of Key Laboratory of Magnetic Suspension Technology and Maglev Vehicle, Ministry of Education.

References

1. Wang, S.Y., et al.: Man-loading high-temperature superconducting Maglev test vehicle. IEEE Trans. Appl. Supercond. **13**(2), 2134–2137 (2003)
2. Kusada, S., et al.: The project overview of the HTS magnet for superconducting Maglev. IEEE Trans. Appl. Supercond. **17**(2), 2111–2116 (2007)
3. Yang, W.J., Qiu, M., Liu, Y., Wen, Z., Duan, Y., Chen, X.D.: Levitation characteristics in an HTS Maglev launch assist test vehicle. Supercond. Sci. Technol. **20**, 281–286 (2007)
4. Beyer, C., et al.: Guideway and turnout switch for the SupraTrans project. J. Phys. **4**, 991–994 (2006)
5. Wang, J.S., et al.: Recent development of high temperature superconducting Maglev system in China. IEEE Trans. Appl. Supercond. **19**(3), 2142–2147 (2009)

6. Guo, F., et al.: Structural parameter optimization design for Halbach permanent Maglev rail. Physica C **470**(3), 1787–1790 (2010)
7. Ma, K.B., McMichael, C.K., Lamb, M.A., Chu, W.K.: Application of high temperature superconductors on levitation bearings, torque transmissions and vibration dampers. IEEE Trans. Appl. Supercond. **3**(1), 388 (1993)
8. Terentiev, A.N., Kuznetsov, A.A.: Drift of levitated YBCO superconductor induced by both a variable magnetic-field and a vibration. Phys. C **195**(1–2), 41 (1992)
9. Terentiev, A.N., Lee, H.J., Kim, C.J., Hong, G.W.: Identification of magnet and superconductor contributions to the ac loss in a magnet-superconductor levitation system. Physica C **290**(3–4), 291 (1997)
10. Lei, W.Y., Qian, N., Zheng, J., Jin, L.W., Zhang, Y., Deng, Z.G.: Dynamic motion modes of high temperature superconducting Maglev on a 45-m long ring test line. Int. J. Mod. Phys. B **31**(25), 1745021 (2017)
11. Wang, B., Zheng, J., Che, T., Zheng, B.T., Si, S.S., Deng, Z.G.: Dynamic response characteristics of high temperature superconducting Maglev systems: comparison between Halbach-type and normal permanent magnet guideways. Physica C **519**, 147–152 (2015)
12. Zhao, L., Deng, J., Li, L., Ning Feng, P., Wei, W.L., Jiang, J., Wang, X., Zhang, Y., Zhao, Y.: The influence of inhomogeneous magnetic field over a NdFeB guideway on levitation force of the HTS bulk Maglev system. Physica C **547**, 41–45 (2018)
13. Chang, P.-Z.: Mechanics of superconducting magnetic bearing. Ph.D. thesis, Cornell University (1991)
14. Teshima, H., Tanaka, M., Miyamoto, K., Nohguchi, K., Hinata, K.: Vibrational properties in superconducting levitation using melt-processed YBaCuO bulk superconductors. Physica C **256**, 142–148 (1996)
15. Qiu, M., Huo, H.K., Xu, Z., et al.: Electromagnetic phenomena in HTS bulk subjected to a rotating field. IEEE Trans. Appl. Supercond. **14**(2), 1898–1901 (2004)
16. Yang, Z.J., Hull, J.R.: Energy loss in superconducting bearing systems. IEEE Trans. Appl. Supercond. **7**(2), 318–321 (1997)

Study on Suspension Stability of High-Speed HTS Maglev System in Evacuated Tube

Dajin Zhou[1,2], Lifeng Zhao[2], Ye Yang[3], Yong Zhang[2], and Yong Zhao[1,2](✉)

[1] College of Physics and Energy, Fujian Normal University, Fuzhou 350117, Fujian, China
yzhao@home.swjtu.edu.cn
[2] Key Laboratory of Magnetic Levitation Technologies and Maglev Trains, Ministry of Education of China, Superconductivity and New Energy R&D Center, Southwest Jiaotong University, Chengdu 610031, China
{zhaolf,yzhao}@home.swjtu.edu.cn
[3] Central Academy Dongfang Electric Corporation, Chengdu 611731, China

Abstract. For the high-speed high-T_c superconducting (HTS) maglev system in evacuated pipe, the side-mounted permanent magnet double-track is used to provide suspension and guidance for the train. When the maglev is running in a round track or a curve, the centrifugal force exerts a side pressure on the track, causing displacement in the suspension direction. In this paper, the static simulation experiment is used to study the instability caused by suspension displacement during the dynamic process of maglev. The results show that under the premise of a given field cooling airgap, the suspension force decreases with the decrease of the suspended airgap and strengthens with the increase of the suspension displacement. In addition, reducing the field cooling airgap can increase the suspension force and reduce the suspension displacement. When the field cooling air gap on the spot is 10 mm and the initial suspension displacement is 2 mm, the suspension displacement will not exceed 2 mm when the train runs to the rated airgap of 5 mm, so that the train maintains high side suspension stability. The research results of this paper provide useful design reference for the application of high speed HTS maglev train in the evacuated tube.

Keywords: Stability · Side-suspension · High-T_c superconductor Maglev · Evacuated tube

1 Introduction

Evacuated tube transportation (ETT) is a new type of transportation that combines vacuum piping technology and maglev train technology, providing the ultimate solution to break the speed limit of traditional ground transportation. At present, the research on evacuated tube transportation mostly stays in the conceptual design or low-speed experimental research stage. In 2013, American technology madman Tesla CEO Elon Musk proposed an evacuated tube transportation program, i.e., the Hyperloop super high-speed rail system. [1] It is designed to be more than twice the speed of the

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 80–86, 2019.
https://doi.org/10.1007/978-3-030-04585-2_10

aircraft. However, to date, the progress of the project has been very limited. In 2015, the magnetic floating loop test speed developed by Deng et al. [2] reached 25 km/h. Huge centrifugal force is the main reason for limiting its high-speed operation.

From 2006 to 2010, Yong Zhao's group [3–6] designed and built the world's first evacuated tube HTS maglev experimental system, and carried out in-depth theoretical and experimental research on the evacuated tube transportation system. The system was designed based on the suspended magnetic track with an auxiliary centripetal force, but the operating speed is still limited by the orbital radius. In 2010-2015, for the research of high-speed evacuated tube transportation system, Zhao et al [7–11] designed and built the second-generation high-speed evacuated tube HTS magnetic levitation experimental system, the train operation mode was changed from up-and-down to the side suspension. In this new system, the permanent magnet track is fixed on the inner side wall of the evacuated tube, so that the high-speed train can obtain greater centripetal force during the movement, thus ensuring that the train does not derail due to the huge centrifugal force.

However, little research has been done on the suspension of this side-hanging form. Preliminary studies have shown [12] that the stability of this side-hung suspension system is challenged by certain factors. As a maglev is running in a round track or a curve, the centrifugal force exerts a side pressure on the track, causing displacement in the suspension direction, which may consequently lead to the instability of the motion for the maglev train. In this paper, the static simulation experiment is used to study the instability caused by suspension displacement during the dynamic process of maglev. The results show that under the premise of a given field cooling airgap, the suspension force decreases with the decrease of the suspended airgap and strengthens with the increase of the suspension displacement. In addition, reducing the field cooling airgap can increase the suspension force and reduce the suspension displacement. When the field cooling air gap on the spot is 10 mm and the initial suspension displacement is 2 mm, the suspension displacement will not exceed 2 mm when the train runs to the rated airgap of 5 mm, so that the train maintains high side suspension stability. The research results of this paper provide useful design reference for the application of high speed HTS maglev train in the evacuated tube.

2 Principle and Design of the Side-Suspension HTS Maglev System in Evacuated Tube

As shown in Fig. 1, the second-generation evacuated tube maglev system is mainly composed of a vacuum pipe, a circular double-track, a ring-shaped long primary linear motor and a base. In Fig. 1(a), when the train is set with a certain floating airgap during the field cooling, it can be stably and freely suspended beside the permanent magnet double rail and directly driven by the long primary coil of the linear motor located directly below it. The structure overcomes the large centrifugal force of the up-and-down suspension loops at high speed, thereby significantly reducing the possibility of train derailment.

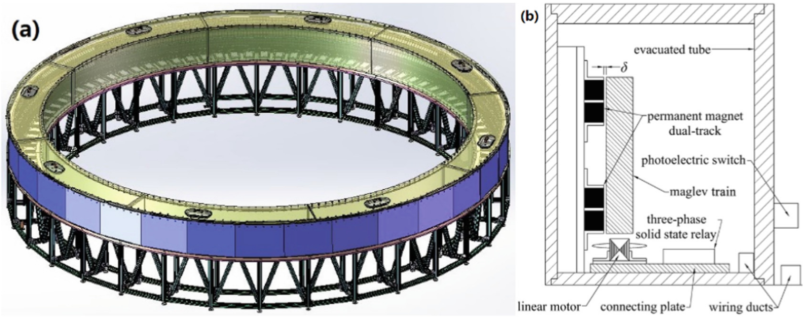


Fig. 1. (a) 3D structure drawing, and (b) 2D sectional drawing for the round track evacuated pipe system.

During the experiment on the 1/8 test section of the evacuated tube (see Fig. 2), after the maglev vehicle is positioned beside the railways with the field-cooling mode, the maglev will not hang in the initial fixed position, but will resettle to a new stable position after moving down a certain distance in the direction of gravity. There is still a small change in the suspension displacement during the operation of the vehicle body. Due to the limitation of the stroke of the test section, we can only study the suspension displacement of the maglev in a low speed operation. In order to study the suspension stability of the high-speed maglev, we used the lateral pressure process of the simulated train at high speed to study the suspension displacement of the evacuated tube maglev system.

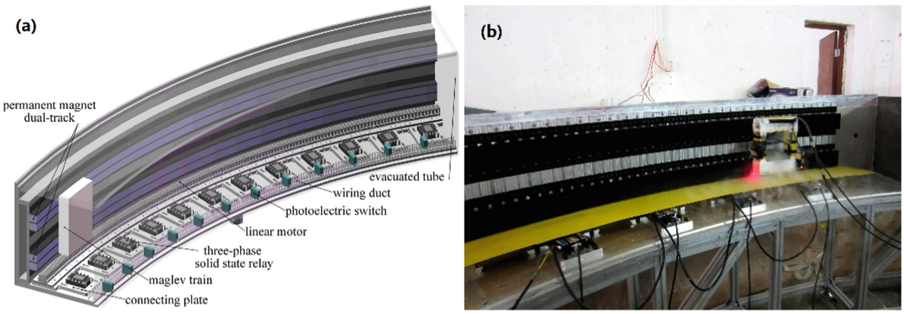


Fig. 2. The design drawing (a) and real construction (b) for a 1/8 test section of the round track evacuated pipe system.

The schematic diagram of the suspension system is shown in Fig. 3(a). Two single-peak permanent magnet tracks are used to provide suspension and guidance for the train. The superconducting maglev vehicle was constructed with a double dewar structure, in which two single dewars are connected by conduits. The superconducting bulks are congested in the dewers with a single-row arrangement to the magnetic peaks of the guideway, for which the highest utilization of suspension force and suspension force can be fulfilled. Different from the conventional up-and-down suspension method, the vehicle body weight in the side suspension system is provided by the guidance force, and the centrifugal force is overcome by the suspension force. This dynamic structure reduces the possibility of derailment of the vehicle at high speed, thereby increasing the running speed of the train. In the actual evacuated tube operation, it is difficult for us to effectively measure the suspension force and guidance force of the train. For this reason, an experimental setup shown in Fig. 3(b) is developed which can control and set the amount of the displacement in the suspension direction, and thus can simulate the effect of centrifugal force caused by the tangent running of the train. The force and displacement are measured by the normal and lateral sensors to simulate the dynamic operation of the maglev.

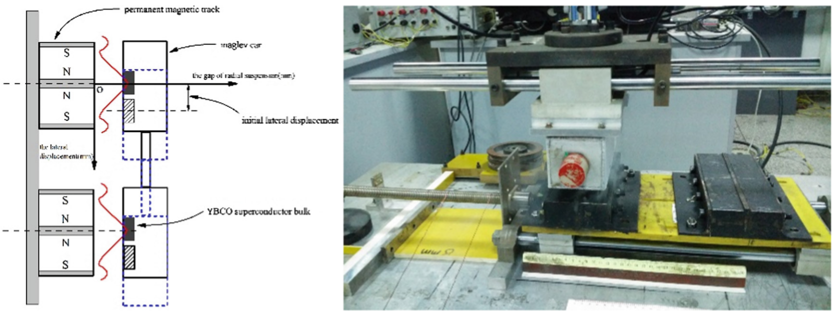


Fig. 3. Schematic (left figure) and experimental (right figure) diagrams for the single magnetic peak double rail side suspension system.

3 Results and Discussion

When static, the maglev vehicle completes the initial positioning in the field cooling mode. When the superconductor bulks enter the superconducting state, the train enters a free-suspending state as the initial positioning support is removed. Due to the presence of gravity, the free-suspending maglev vehicle will generate an initial displacement in the direction of gravity, which induces a suspension force to overcome the weight of the maglev vehicle and allow the vehicle to enter a stable, completely free-hanging state. When the maglev vehicle runs at high speed along the curve, the centrifugal force received by the maglev vehicle will increase with the square of the running speed. Under the action of this centrifugal force, the maglev vehicle gradually approaches the magnetic track, and the suspension airgap is reduced. By means of the simulation experiment set-up, we studied that at different field cooling heights, when the relative displacement of the fixed track and the maglev vehicle in the longitudinal direction is constant, and the train is gradually orbiting, we find that the suspension force follows the suspended airgap, as shown in Fig. 4(a). When the suspension force is reduced to less than the initial suspension force (vehicle weight), a new suspension displacement will inevitably occur due to excessive vehicle weight (or overload). As shown in Fig. 4(b), on the basis of the previous experiment, when the suspension force is reduced, the relative displacement of the maglev vehicle and the track is increased, and it is found that as the airgap in field cooling is small, the suspension force is not reduced, but the suspension force at the rated working point is always greater than the initial suspension force (vehicle weight). In contrast, as the airgap in field cooling is large, the curve between airgap and suspension force shows a parabola with a maximum value. By comparing the two figures in Fig. 4, when the vehicle body produces a suspension displacement, the suspension force can be effectively suppressed. The reduction is more prominent in the small field cooling airgap. Under the condition of a certain suspension displacement, the field cooling airgap exhibits a stable airgap range.

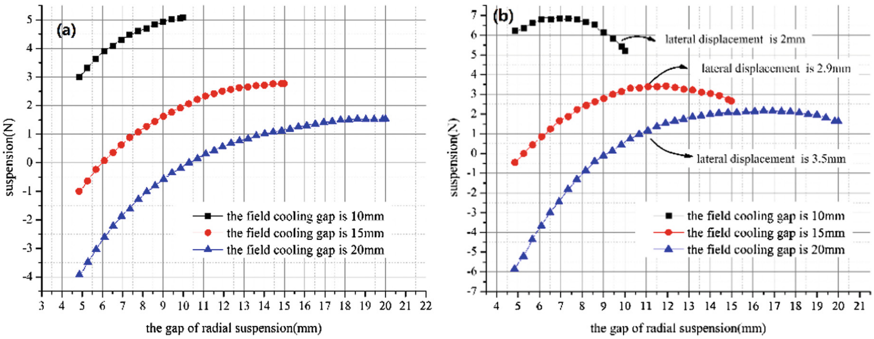


Fig. 4. Relationship between suspension force and airgap (a) with suspension displacement; and (b) without suspension displacement

In order to further study the characteristics of the airgap-suspension force, when the maglev is close to the track by the side pressure, the suspension force is always equal to the initial suspension force (vehicle weight) by controlling the displacement of the vehicle. When the vehicle runs to the rated working airgap point, the dynamic stability characteristic curve of the airgap-suspension displacement can be obtained, as shown in Fig. 5(a). In the case where the initial displacement of the vehicle body suspension is 2 mm, in order to stabilize the vehicle operation, the amount of suspension displacement generated during the operation must not exceed 2 mm. For the case of large initial airgaps (20 mm and 15 mm), when the running airgap reaches 12 mm and 8 mm respectively, the suspension displacement will reach 2 mm. If the airgap is further reduced, it is necessary to generate a larger suspension displacement to balance the loss of suspension force due to the reduction of the airgap. Based on this, we can explain why the airgap-suspension parabola characteristic curve with maximum value will appear in the lateral pressure process when the airgap field is large. This is because the airgap is so larger, and a larger suspension displacement is required during the lateral pressure to maintain the stability of the suspension. If the amount of displacement generated is insufficient to compensate for the loss of the suspension force, the suspension force will be less than the initial suspension force (vehicle weight) and instable phenomenon occurs, as shown in Fig. 5. In the case of low airgap (10 mm), the suspension force during the whole side pressure operation does not appear less than the initial suspension force, and the lateral displacement generated when the vehicle runs to the rated working airgap keeps within 2 mm (Fig. 6).

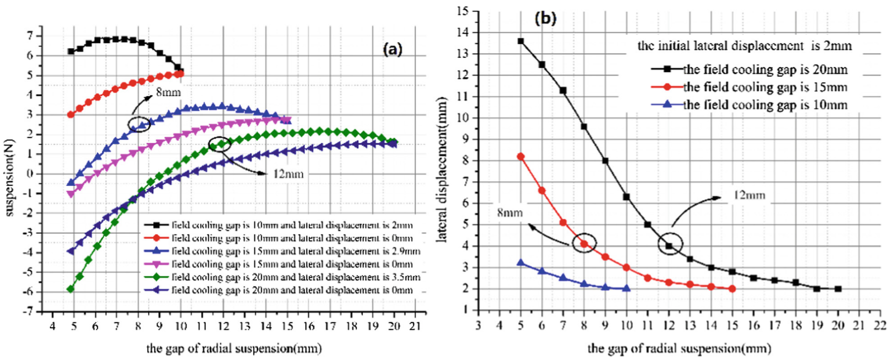


Fig. 5. (a) Relationship between suspension force and airgap under various initial airgaps; (b) Relationship between lateral displacement and airgap under various initial airgaps.

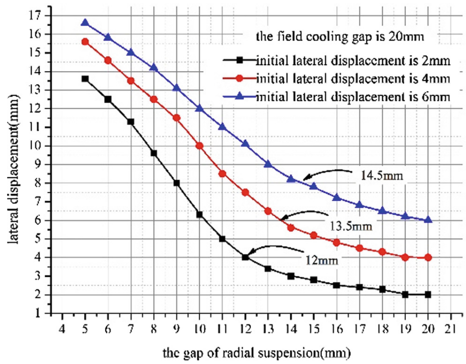


Fig. 6. Relationship between lateral displacement and airgap under various initial airgaps generated by different vehicle weights.

4 Conclusion

In summary, the instability caused by suspension displacement during the dynamic process of maglev in studied by a static simulation experiment in this paper. It is observed that under the premise of a given field cooling airgap, the suspension force decreases with the decrease of the suspended airgap and strengthens with the increase of the suspension displacement. In addition, the suspension force is increasing with reducing the field cooling airgap, but the suspension displacement is decreased at the same time. When the field cooling air gap on the spot is 10 mm and the initial suspension displacement is 2 mm, the suspension displacement will not exceed 2 mm when the train runs to the rated airgap of 5 mm, so that the train maintains high side suspension stability. The research results of this paper provide useful design reference for the application of high speed HTS maglev train in the evacuated tube.

Acknowledgements. The authors are grateful to the financial support of the Sichuan Applied Basic Research Project (No. 2018JY0003, 2017JY0057), the Science and Technology Research and Development Program of Sichuan Province (No. 2016GZ0147), and the National High-Tech R&D Program of China (863Program) under Grant 2014AA032701.

References

1. http://tech.qq.com/a/20170712/068300.htm?t=1499894479308
2. Deng, Z., Zhang, W., Zheng, J.: A high temperature superconducting Maglev ring test line developed in Chengdu, China. IEEE Trans. Appl. Supercond. **26**(6), 1–8 (2016)
3. Ma, J., Zhou, D., Zhao, L., et al.: The approach to calculate the aerodynamic drag of maglev train in the evacuated tube. J. Mod. Transp. **21**(3), 200–208 (2013)
4. Chen, X., Zhao, L., Ma, J., Liu, Y.S.: Aerodynamic simulation of evacuated tube maglev trains with different streamlined designs. J. Mod. Transp. **20**(2), 115–120 (2012)
5. Ma, J., Zhou, D., Zhao, L., Zhang, Y., Zhao, Y.: Vibration energy consumption characteristics of high temperature superconducting magnetic levitation vehicle in vacuum pipeline. J. Vac. Sci. Technol. **34**(2), 119–125 (2014)
6. Ma, J., Zhou, D., Zhao, L., Zhang, Y., Zhao, Y.: Electromagnetic brake analysis of vibration energy dissipation method in vacuum pipe HTS magnetic float system. J. Vac. Sci. Technol. **35**(2), 130–136 (2015)
7. Zhou, D., Cui, C., Zhao, L., Zhang, Y., Wang, X., Zhao, Y.: Static and dynamic stability of the guidance force in a side-suspended HTS maglev system. Supercond. Sci. Technol. **30**(2), 025019 (2017)
8. Zhou, D., Cui, C., Zhao, L., Zhang, Y., Wang, X., Zhao, Y.: Running stability of a prototype vehicle in a side-suspended HTS maglev circular test track system. IEEE Trans. Appl. Supercond. **27**(1), 3600107 (2017)
9. Zhou, D., Ma, J., Zhao, L., Zhang, Y., Guo, J., Zhao, Y.: Vacuum tube HTS maglev train experimental system ring accelerator design. J. Vac. Sci. Technol. **35**(4), 391–398 (2015)
10. Zhou, D., Cui, C., Zhao, L., Zhang, Y., Guo, Z.: Y.: Research on starting characteristics of linear motor in vacuum piping HTS side floating system. J. Southwest Jiaotong Univ. **51**(4), 750–758 (2016)
11. Zhou, D., Zhao, L., Ke, C., Hsieh, C.C., Cui, C., Zhang, Y., Zhao, Y.: High-Tc superconducting maglev prototype vehicle running at 160 km/h in an evacuated circular track. IEEE Trans. Appl. Supercond. **28**(4), 3600504 (2018)
12. Liang, G., Zhao, L., Yang, J., Ma, J., Zhang, Y., Wang, X., Zhao, Y.: Study of the maglev performance of the side-mounted high-temperature superconductor maglev rotating system. IEEE Trans. Appl. Supercond. **25**(4), 3601406 (2015)

Design and Realization of Intelligent Voice-Control Car Based on Raspberry Pi

Lisang Liu[1,2](✉), Shaoping Zhu[1], Dongwei He[1,3], Ying Ma[1,3], Xiuzhen Zhang[1], Jing Huang[1,3], and Junnan Li[1]

[1] Technical Development Base of Industrial Integration Automation of Fujian Province, Fujian University of Technology, Fuzhou 350118, China
liulisang@fjut.edu.cn
[2] FuJian Key Laboratory of A.E.D, Fujian University of Technology, Fuzhou 350118, China
[3] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education, Fujian University of Technology, Fuzhou 350118, China

Abstract. This paper introduces an intelligent voice-control car based on Raspberry Pi. The user's speech is recorded and uploaded to Baidu cloud to be recognized and analyzed by Baidu speech recognition technology. The intelligent car movement can be controlled according to the returned recognition results if control commands included. Meanwhile, Turing robot technology enables the car to dialogue with the users. The car can also be remotely controlled by using computers and other intelligent devices in real time. The testing shows that no matter it is on-site operation or remote operation, the car can work efficiently with high correction of speech recognition.

Keywords: Intelligent car · Speech recognition · Remote control Turing

1 Introduction

Convenient and efficient human-machine interaction is required as the intelligent devices emerge in an endless stream. Speech is one of the best ways of interaction between intelligent devices and users, even the children and the old people can use it [1]. In the past two decades, many effective speech processing and recognition methods and strategies have been put forward, which makes the research of speech recognition become more and more prosperous. Since 1990s, many famous big companies such as IBM, apple, AT&T and NTT have invested heavily in the practical research of speech recognition system [2–4]. While in China, the speech recognition research has developed to large vocabulary, non-specific person and continuous speech, and has received much attention from government and fund departments. In April 1998, the national 863 intelligent computer expert group evaluated the domestic large vocabulary continuous speech recognition systems, and the system of the Department of electronic engineering of Tsinghua University had obtained the best results with the 93% correct rate of the word and the 62.5% correct sentence rate. This result was equal to the level of the IBM

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 87–95, 2019.
https://doi.org/10.1007/978-3-030-04585-2_11

speech recognition system. The accuracy of the continuous speech dictation system and Chinese human-machine dialogue system, which were developed by Chinese Academy of Sciences Automation Institute, was over 90% [5]. There are some other systems with high accuracy of speech recognition, such as IBM's ViaVoice and Dragon Naturally Speaking System, Nuance's Nuance Voice Platform, Microsoft Whisper, Sun Voice Tone and so on. The correct rate of Google speech recognition can reach more than 95% [6], while the rate is more than 97% among the three voice recognition magnates of Baidu, IFLYTEK and Sogou.

At present, speech recognition has been widely used, one of the important applications is the vehicle voice system in intelligent car, which enables drivers to control the car quickly and conveniently through the language, such as listening to music, sending messages, calling, seeking help, navigation, auxiliary driving and so on.

Speech recognition technology is a high-tech technology that allows the machine to convert speech signals into corresponding text or commands through recognition and understanding process. It has been widely used in the field of intelligent vehicles. This paper designs a voice-control car based on Baidu cloud speech recognition technology, which has the functions of voice control, voice dialogue, remote control and so on. It has certain application value.

2 System Scheme

To design an intelligent voice-control car, the key problems are as follows:

(1) Speech recording and speech recognition.

The user's voice information is recorded and speech recognition is completed after the recording is completed, then the corresponding text after identification is returned.

(2) Communication with the intelligent voice-control car.

The voice information is communicated with the intelligent voice car. The intelligent voice-control car receives speech recognition data, and dialogues according to its content.

(3) Playing out the content of the dialogue in audio.

The text returned from the Turing robot is synthesized into the voice information through speech synthesis, then played out so that users can hear it.

(4) Control the movement of the car according to the speech recognition

If the voice information contains control instructions, the intelligent voice-control car should move in accordance with instructions.

(5) Remote control of the car movement.

Not only can we control the intelligent car on the spot, but also we can control the movement of the intelligent car remotely by using a computer or smart phone.

Thus, we proposed a set of scheme for intelligent voice-control car, shown in Fig. 1.

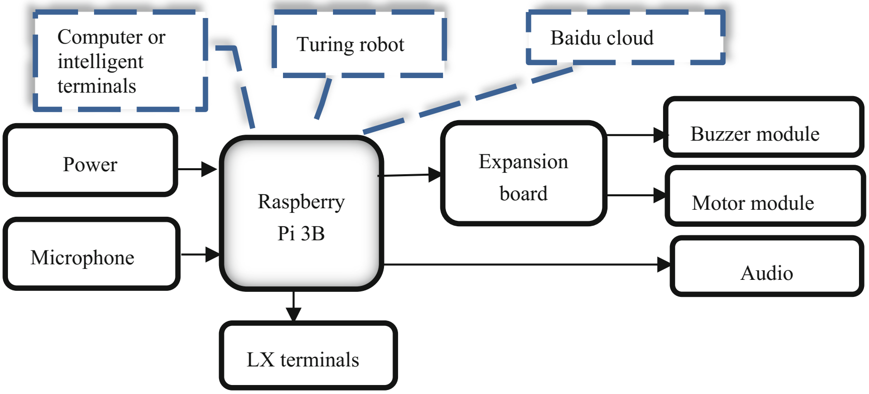


Fig. 1. Scheme of intelligent voice-control car

In our design scheme, the Raspberry Pi 3B is adopted as the control unit. The Raspberry Pi 3B is a development board with CPU and memory, whose CPU performance is far beyond the single chip [7]. The Linux operating system can be run on it. The Raspberry pie 3B is powerful because it adds a lot of expanded peripherals, and integrates TB6612FNG modules and active buzzer modules on the expansion board. The TB6612FNG module, manufactured by Toshiba Corporation Semiconductor, is used for driving the DC motors. The module adopts two way driving, one way controls two left wheels and the other controls two right wheels.

2.1 Voice Control Scheme

Since 2017, Baidu, IBM, Microsoft and HKUST have adopted Convolutional Neural Network (CNN) as the framework of speech recognition. CNN is one of the hottest directions of speech recognition in the last two years due to its invariance in the frequency domain and the proposal of deep CNN networks such as VGG and residual network, which bring new development to CNN. With the developed end to end CTC framework and more than 10 layers deep network structure, companies have made remarkable achievements in the application of deep CNN [8]. Baidu, one of the famous companies, can achieve high correct speech recognition rate by uses deep CNN network structure and VGGNet framework. The deep CNN structure containing Residual connection, and combines the end to end speech recognition technology of LSTM and CTC.

With the help of Baidu mature speech recognition technology, the user's voice, which is sent to the cloud, is recognized and the recognized data is returned to the local area from the cloud. The returned data is processed by the specific program on the local machine. The intelligent car performs the corresponding action if control commands are included. Otherwise, the Turing robot talks with the user to achieve human-computer interaction. In other words, no matter what speech information the user enters, the car can always respond accordingly, such as inquiring about the weather, chatting, checking train schedules and so on.

2.2 Remote Control Scheme

Considering some special applications, remote login and control of the intelligent car is necessary. That is to say, while the program is running, a computer or a smart phone connected to the same LAN can control the car by keyboard inputting or logging on a web page and clicking on the specific area. The advantage of the remote control is that the car can be controlled by using other intelligent devices in real time regardless of sheltered objects and far distances, as long as they are on the same LAN. When writing a wireless control program, one should pay attention to determining whether the keycode of the keyboard is a valid key or whether the area clicked by the mouse is a valid area for the car action.

3 Hardware Implementation

The four corners of Raspberry Pi 3B can be fixed to the board of the intelligent car by screws. Taking heat dissipation into account, four copper columns are used to keep a certain height between the Raspberry Pi and the board, so as to avoid direct contact of the two and short circuit problems. After the Raspberry Pi is installed, the extension board is connected to the Raspberry Pi through female header. Then screw it on the Raspberry Pi.

Put the specific free-drive microphone in the Raspberry Pi and test the recording effect of the microphone by entering the following code at the LX terminal: *sudo arecord -D "plughw:1" -f S16_LE -r 16000 -d 3 /home/pi/Desktop/voice.wav*, then the sound is played out by audio.

The four DC decelerating motors, which are connected with the four wheels of the intelligent car respectively, are fixed on the chassis with screws and nuts. The battery box is also fixed under the chassis (Figs. 2, 3 and 4).

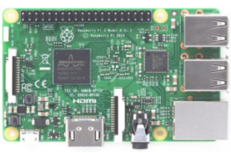


Fig. 2. Raspberry Pi 3B

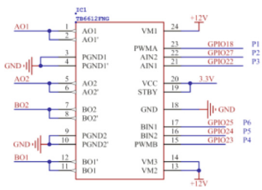


Fig. 3. Wiring of TB6612FNG

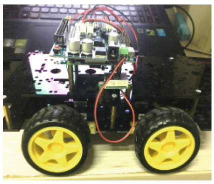


Fig. 4. The installed car

4 Software Programming

In order to make the car intelligently working, the system should be well designed. The main idea of voice control program is, firstly, to detect the sound after the program starts running, the voice information that the user input will be recorded at the detection. Second, after judging the end of the speech, the system uploads the recording file to the Baidu cloud to recognize and the recognition results by the Baidu speech recognition technology will be returned. Third, if the control command is contained in the recognition result, the intelligent car will perform the corresponding action. If there is no control command, the Turing robot will dialogues with the user with a returned text information, which as well be synthesized through the speech synthesis program and be played out by the audio (Fig. 5).

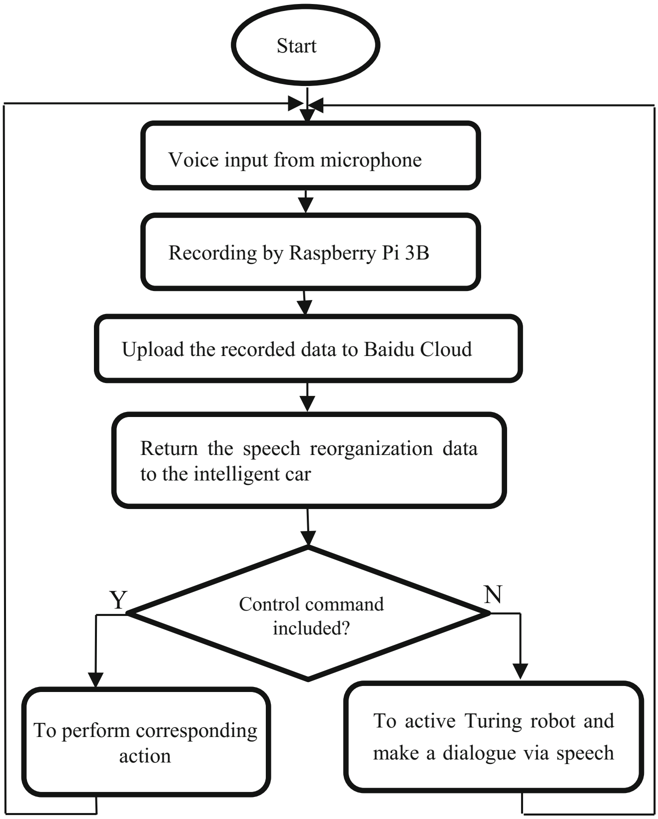


Fig. 5. Voice control program

The specific steps are as follows.

(1) Register an AI account at http://ai.baidu.com and create an application (Table 1).

Table 1. Apply for an AI secret key.

App ID.	API key	Secret key
11102644	rYDhHNFuxwAdlbl4C5KVyHQC	ybmxZ0e1mjfyyL1Ub2cghlcLCQppMo0u

After registration, you will get APPID, APIKEY, and Secret Key as shown above, which can be used to call speech recognition and voice synthesis, based on the methods of invoking Python and Baidu AI.

(2) Registered Turing robot at http://www.tuling123.com, then create a robot (Table 2).

Table 2. Apply for turing robot secret key.

API key	Secret key
acb27dc18242442e9913ff9a2079eadf	1a2a719925afc8c7

After registration, you can get the APIKEY and Secret Key as shown in the table above, which can be used to call the Turing robot, according to Python and Turing robot call methods.

5 Simulation and Testing

5.1 Speech Testing

The voice control commands including *moving forward, backward, turn left, turn right, speed up/accelerate, slow down/decelerate, stop, moving for three seconds* and so on. When these kinds of commands are recognized, the intelligent car performs the corresponding action as it is told. While non-control commands are recognized, the intelligent car stay still or keep the original state of the motion, the Turing robot will analyze the recognized information and make a corresponding reply so as to realize the human-computer interaction.

First, when program *yuyin.py* is running, the LX terminal shows the recording instruction. The users can input control commands by microphone. The initial running interface is as follows (Fig. 6).

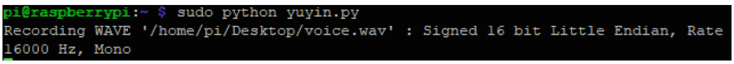


Fig. 6. Initial interface on LX terminal

When a control command is issued, such as moving forward, the LX terminal will display the corresponding command, and the car will keep moving forward until the next control command is issued. The system keeps recording in the course of car movement.

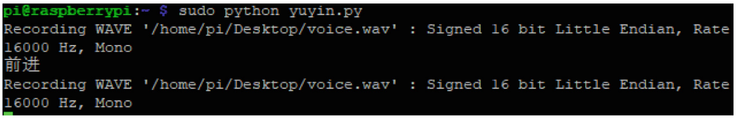


Fig. 7. Speech testing 1

When a non-control command is issued, for example, 'How is the weather in Bayan Nur?' The car will tell the local weather with voice and the LX terminal will show the weather information, too (Fig. 8).

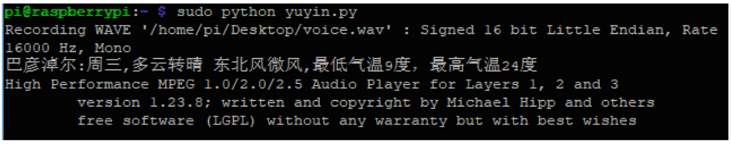


Fig. 8. Speech testing 2

We also preset up some interesting answers in the car program to make the dialogue funnier. For example, when the intelligent car was asked 'who are you'? It will reply that 'just call me Steamed Buns' (Fig. 9).

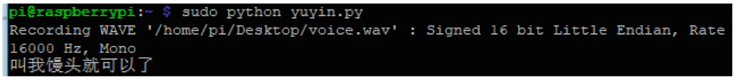


Fig. 9. Speech testing 3

5.2 Remote Control Testing

The remote control testing includes two parts, one is tested by remote computer within the same LAN, and the other is tested by intelligent terminals such as smart phone.

When program *index.py* is running, enter the IP address of the Raspberry Pi in browsers on computer or smart phone, the control interface of the intelligent car will appear on the webpage as shown below. The IP address is *172.20.10.3*. Click the valid square area on a smartphone or other intelligent terminals, the car will perform its action of moving forward, turn left, moving backward, turn right, decelerate, accelerate, stop and whistle, which are corresponding to the keys W, A, S, D, Q, E, Z and X on the keyboard respectively (Fig. 10).

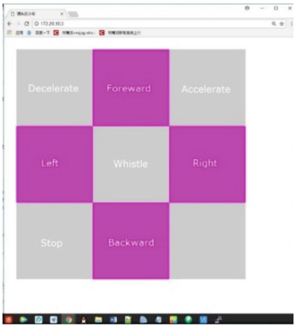


Fig. 10. Control interface on the webpage

We tested the W key by entering the IP address on the computer, for example, the intelligent car moved on and the LX terminal displayed ‘moving forward’, the same as the Fig. 7 showed. Then we tested again by click the colored squares on smart phone and other intelligent terminals, take the ‘left’ square for instance, the intelligent car turned left immediately. When ‘decelerate’ or ‘accelerate’ was clicked, the car slowed down or speeded up at a fixed rate until it reached the minimum or maximum velocity. Meanwhile, the motion state and the present velocity were displayed on the LX terminal at the same time.

6 Conclusion

Speech is an efficient and fast way of interaction, thus intelligent car with speech recognition and control has wide application fields. In this paper, the speech information is collected through the microphone, and the Raspberry Pi is adopted as the control core. The Baidu cloud speech recognition technology is used to carry out speech recognition and analysis. The voice control, dialogue communication and remote control of the car can be carried out according to the content of the speech information. It can not only control the direction and speed of the car, but also control the driving time of the car. The remote control module enables users to interact with the car on the intelligent terminals or computers through LAN. In addition, the car also has whistle alarm function, emergency braking and other functions.

To make a summary, the intelligent voice-control car implemented in this paper combines Baidu AI, Turing robot, python, JavaScript, remote control and other technologies based on Raspberry Pi and its expansion board, with rich and extensible functions in speech query and control.

Acknowledgement. The authors would like to thank the anonymous reviewers for their valuable comments. This work was supported by Initial Scientific Research Fund of FJUT (GY-Z12079), Pre-research Fund of FJUT (GY-Z13018), Fujian Provincial Education Department Youth Fund (JAT170367,JAT170369), Natural Science Foundation of Fujian Province (2018J01640) and China Scholarship Council(201709360002).

References

1. Yi, K., Tian, B., Fu, Q.: Design and realization of vehicle speech recognition and control system. Comput. Appl. Softw. **34**(5), 129–134 (2017)
2. Yang, H., Chi, H., et al.: Speech Signal Digital Processing. Electronics Industry Press (1995)
3. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Tsinghua University Press (1999)
4. Rabiner, L.R.: A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE **71**(2), 257–285 (1989)
5. Dai, S.: Design of Vehicle Speech Recognition and Control System. Chongqing University of Posts and Telecommunications (2011)
6. Pan, Y.: Research on Speech Recognition System Based on Improved HMM/SVM Architecture. Lanzhou University (2009)
7. Raspberry Pi Hardware [EB/OL]. https://www.raspberrypi.org/documentation/hardware/
8. Application of convolution neural network (CNN) in speech recognition [EB/OL]. https://www.cnblogs.com/qcloud1001/p/7941158.html?utm_source=debugrun&utm_medium=referral. Accessed 01 Dec 2017

LSTM Power Mid-Term Power Load Forecasting with Meteorological Factors

Xin Su[1,3], Xin-hua Jiang[1(✉)], Shun-miao Zhang[1,2], and Ming-long Chen[1,2]

[1] College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350108, Fujian, China
xhjiang@fjut.edu.cn

[2] Fujian Provincial Key Laboratory of Big Data Mining and Applications, College of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, Fujian, China
zshunmiao@163.com

[3] Fujian Key Laboratory for Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350118, Fujian, China
xjlw66@163.com

Abstract. In order to improve the accuracy and efficiency of mid-term power load forecasting, a mid-term power load forecasting method of long short time memory network (LSTM), which combines weather factors, is proposed. Firstly, the influence of meteorological factors affecting the power load on the mid-time power load is analyzed. Secondly, the meteorological factors are used as the input factor of the LSTM model to predict the power load. Finally, compared with the Random forest, ARIMA, GBDT, the LSTM algorithm which is not fused with meteorological factors, through the analysis of the experimental data, the LSTM fusion of meteorological factors has a better prediction effect on the mid-term load forecasting.

Keywords: Mid-term power load forecasting · Meteorological factors Long and short time memory network

1 Introduction

The mid-term power load forecasting is the basis of the power system operation and analysis, and is an important part of the EMS system. It is a great significance to the optimization combination of the power unit, the dispatching of the water and the fire and the electricity, the economic distribution of the load. Improving the accuracy of load forecasting is an important way to ensure the scientific decision-making of power system optimization. Scholars at home and abroad have made fruitful research on electric power load forecasting. It is mainly divided into several kinds of methods:

Traditional power load forecasting methods, a non-linear regression model for mid-term load forecasting and improvements in seasonality is proposed in [1].

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 96–103, 2019.
https://doi.org/10.1007/978-3-030-04585-2_12

It presented some approaches to improve the modeling of seasonality. But requires great care in the day-type typology.

The modern load forecasting method, a method to predict probability density of medium-term power load considering temperature factor is proposed in [2]. According to continuous conditional quantile functions the probability density of medium-term power load on a certain day is predicted to obtain more information related to medium-term power load.

Other load forecasting methods, the ideal-point and multi-attribute decision algorithm for medium and long-term power load forecasting is proposed in [3]. The results are compared with the actual peak load data to attain minimum percentage error. From average error point of view, it is found that PSO algorithm have produced better estimation than the AR model.

In conclusion, it is found that the latest forecasting methods pay more and more attention to the changing trend of historical data and the power load forecasting under the influence of multiple factors. Therefore, this paper proposes a mid-term power load forecasting method based on a LSTM fusion of meteorological factors. Firstly, the correlation analysis is carried out on the meteorological factors, and the factors that affect the power load are selected and the meteorological factors are normalized. Then the meteorological data are processed into the LSTM model to predict mid-term power load. Finally, we compare the prediction results of this method with the common mid-time load forecasting method: ARIMA, random forest, and GBDT.

2 Influence Factors of Mid-Term Power Load Forecasting

2.1 Screening of Influence Factors of Power Load

In modern power system, considering meteorological factors has become one of the important means to improve the accuracy of load forecasting. The authors collected 8 Meteorological characteristics, including maximum temperature, minimum temperature, average temperature, relative humidity, rainfall, air pressure, wind direction and wind speed. The Gini coefficient [4] is used to evaluate the importance of features in order to determine the importance of features. The random forest algorithm is used for feature importance screening and drawing. The variable importance measures is expressed by VIM, and the Gini exponent is expressed in GI, assuming that M features $X_1, X_2, \ldots X_m$. calculate the Gini exponent score of each characteristic x_j of VIM_j^{gini} process is as follows:
The formula for the Gini index is:

$$GI_m = 1 - \sum_{k=1}^{|k|} p_{mk}^2 \tag{1}$$

Among them, K indicates that there are K categories, and p_{mk} represents the proportion of category K in node m. The importance of characteristic x_j in node m is that the variation of Gini exponent before and after M branch is:

$$VIM_j^{Gini} = GI_m - GI_l - GI_r \tag{2}$$

Among them, GI_l and GI_r are the Gini exponents of two new nodes respectively. Suppose that there are n trees in the random forest, then:

$$VIM_j^{Gini} = \sum_{i=1}^{n} VIM_{ij}^{Gini} \tag{3}$$

Finally, all importance scores were normalized.

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{m} VIM_i} \tag{4}$$

The importance Association of meteorological features is as Fig. 1:

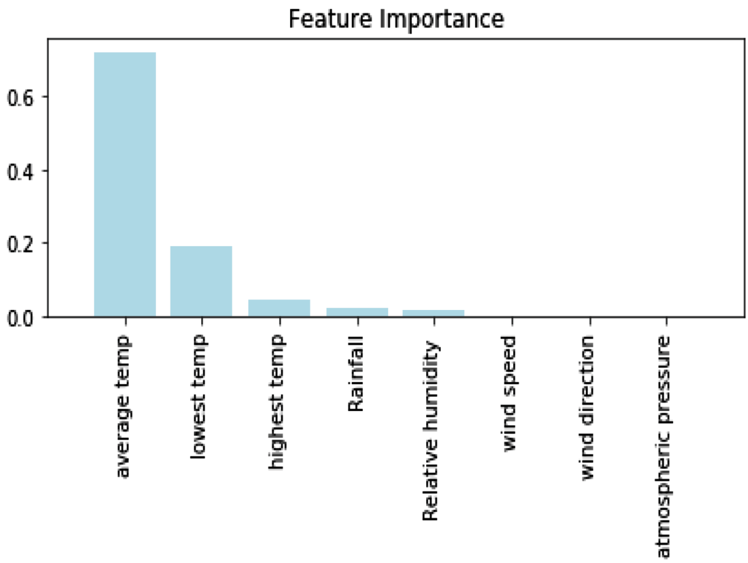


Fig. 1. Feature important

According to the characteristic importance association diagram, the highest temperature, the lowest temperature, the average temperature, the relative humidity and the rainfall are selected as the meteorological characteristics of the main influence. The LSTM model is constructed as a training set to import multi-target long and short time memory networks.

2.2 Normalization/Scaling

The normalized input and output data are used for training of LSTM model. The common normalization approach includes statistical normalization and Min-Max normalization. In this paper, Min-Max normalization is used. The input feature demands are normalized with respect to their lower and upper values in each training pattern using

$$x_{norm} = \frac{x - min(x)}{max(x) - min(x)} \tag{5}$$

Where x is the original value and x_{norm} is the value after regularization. The regularized input data are imported into the LSTM model for training learning.

3 Long and Short Time Memory Network (LSTM [5]) model

The long and short term memory model is a special RNN model, which is proposed to solve the problem of the gradient dispersion of the RNN model. The LSTM structure is shown as follows:
Forget gate:

$$f_t = \sigma(W_f[h_{t-1}, X_t] + b_f) \tag{6}$$

The gate reads h_{t-1} and X_t, and outputs a numeric value between 0 and 1 to each in unit memory C_{t-1}, of which numbers, 1 represent "completely reserved", and 0 means "completely abandoned".
Input gate:

$$i_t = \sigma(W_i[h_{t-1}, X_t] + b_i) \tag{7}$$

$$\tilde{C}_t = tanh(W_C[h_{t-1}, X_t] + b_C) \tag{8}$$

Sigmoid decides what value will be updated. Then, create a new candidate value vector through the tanh layer.
Update gate:

$$\tilde{C}_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$

This layer is used to update the current unit memory C_t
Output gate:

$$O_t = \sigma(W_o[h_{t-1}, X_t] + b_o) \tag{10}$$

$$h_t = O_t * tanh(C_t) \tag{11}$$

The sigmoid layer determines which part of the unit memory will be exported. Processing through tanh (getting a value between −1 and 1) and multiplying it with the output of the sigmoid gate to get the final output.

4 Experiment and Result Analysis

4.1 Data Descriptions

The data set was used to test the mid-term load forecasting data of power system in the 2016 National College Students' Electrical Mathematical Modeling Competition A sponsored by the Electrical Engineering Society. A total of 1114 power load data for 2012–2015 years. Take 2013-2014 years data as training set, and last 48 days in 2015 as test set.

4.2 Performance Metrics

In this paper, the proposed model was evaluated using root mean square error, mean absolute error, and absolute median error. The specific evaluation algorithm is shown as follows:

(1) RMSE (Root mean square error)

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n_{sample}} \Sigma_{i=0}^{n_{sample}-1} (y_i - \hat{y}_i)^2} \tag{12}$$

(2) MAE (Mean absolute error)

$$MAE(y, \hat{y}) = \frac{1}{n_{sample}} \Sigma_{i=0}^{n_{sample}-1} |y_i - \hat{y}_i|^2 \tag{13}$$

(3) MedSE (Absolute median error)

$$MedSE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|) \tag{14}$$

Where y, $\hat{y}$ represent the true value and the predicted value respectively.

4.3 Experimental Steps

First, the data is introduced into the LSTM prediction model, and the first hidden layer of LSTM is defined with 50 neurons, and the output layer is 1 neurons, and the output is the predictive load value. Second, the input shape is a 1 time step with 5 characteristics. The average absolute error (MAE) loss function is used in the LSTM model. finally, the optimizer uses the 'adam' optimizer [6]. The number of iterations is 250. The convergence effect is as Fig. 2.

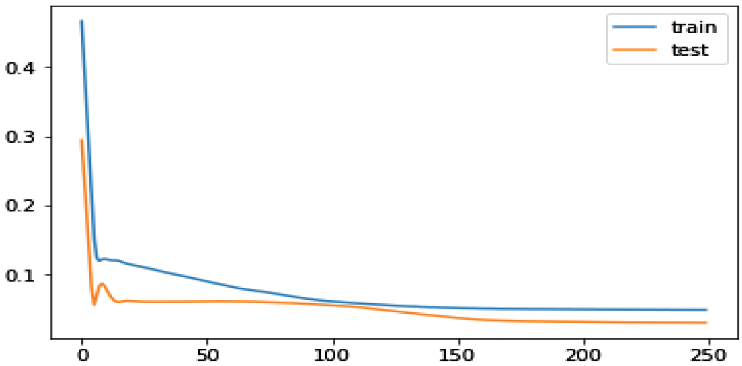


Fig. 2. LSTM network convergent graph

The loss coefficient of the final training set is 0.0490, and the loss coefficient of the test set is 0.0303. The result shows that the LSTM model can converge better.

4.4 Result Analysis

In this paper, the LSTM fusion of meteorological factors is referred to as FM-LSTM. The trained FM-LSTM model was compared with the real load curve and the traditional LSTM model. The results are as Fig. 3.

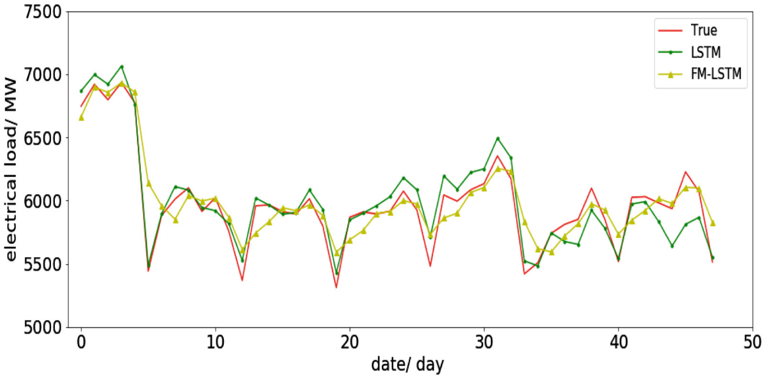


Fig. 3. FM-LSTM network

The root mean square error (RMSE) is 105.44. It can be seen from the Fig. 3 that the LSTM model without considering meteorological factors and FM-LSTM model with considering meteorological factors can better fit the trend of mid-term power load. The LSTM model, which does not consider the meteorological factors, is generally high, and in the treatment of abrupt weather, for example, the 4 day is rainy day, the temperature drops, resulting in the decrease of electricity load. There is a prediction bias at this point without considering the meteorological factors. The LSTM considering meteorological factors can better predict the LSTM. In summary, the FM-LSTM model can better fit the load curve. At the same time, to verify the performance of the model, the proposed method is compared with three common power load forecasting algorithms, namely ARIMA [7], random forest [8], GBDT [9]. The results are as follows:

From the results of the performance comparison between Fig. 4 and Table 1. ARIMA fitted well for periodicity, but the effect on external environmental factors was poor. Among them, GBDT obviously deviates from the predicted target at some points, showing the worst performance. Considering the integration of meteorological factors, the multi-objective LSTM model performs well in nonlinear fitting and periodic fitting.

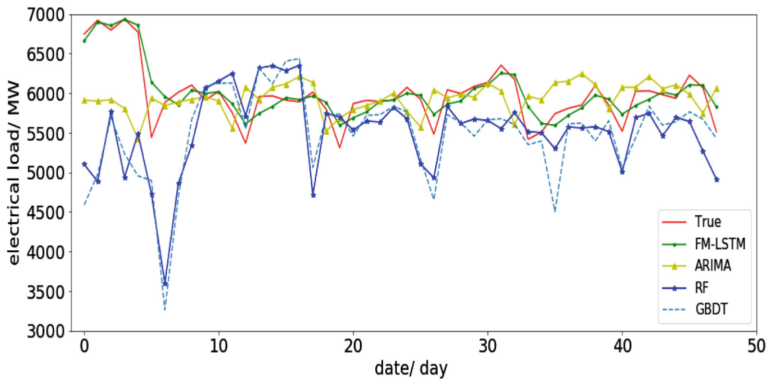


Fig. 4. Compare other method

Table 1. Performance comparison.

Method	Experimental mean (day)		
	RMSE	MAE	MedAE
Random forest	13.3089	10.4917	9.1783
GBDT	16.1360	11.4584	9.0815
ARIMA	11.2670	8.1204	5.7087
Multi-LSTM	**2.6361**	**2.2179**	**2.2575**

5 Conclusion

In this paper, the gini coefficient is used to evaluate the importance of the meteorological features which may affect the mid-term forecasting of power load, and five meteorological features are selected. The long short time memory network model (FM-LSTM), which combines meteorological factors, is used to predict the mid-term power load. The FM-LSTM model performance is better than other methods.

Acknowledgments. This work was supported by the Project of Industry-guidance of Fujian Province under Grant No. 2015H0009, Fujian Young Teacher Education Research Foundation under Grant No. JA14217, National Natural Science Foundation of China (No. 61304199), Fujian Science and Technology Department (No. 2014H0008), Fujian Transportation Department (No. 2015Y0008), Fujian Education Department (No. JK2014033, JA14209, JA1532), and Fujian University of Technology (No. GYZ13125, 61304199, GY-Z160064). Many thanks to the anonymous reviewers, whose insightful comments made this a better paper.

References

1. Bruhns, A., Deurveilher, G., Roy, J.S.: A non linear regression model for mid-term load forecasting and improvements in seasonality (2005)
2. Yaoyao, H.E., Wen, C., Qifa, X.U., Han, A.: A method to predict probability density of medium-term power load considering temperature factor. Power Syst. Technol. **39**, 176–181 (2015)
3. Yang, H., Hui, L.I., Zhan, S., Yan, F.: Medium and long-term power load combination forecasting based on ideal-point and multi-attribute decision algorithm. J. Yanshan Univ. (2017)
4. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. Comput. Sci. (2014)
5. Kong, W., Dong, Z.Y., Jia, Y., Hill, D.J., Xu, Y., Zhang, Y.: Short-term residential load forecasting based on LSTM recurrent neural network. IEEE Trans. Smart Grid (2017)
6. Zhang, X., Pang, Y., Cui, M., Stallones, L., Xiang, H.: Forecasting mortality of road traffic injuries in China using seasonal autoregressive integrated moving average model. Ann. Epidemiol. **25**(2), 101–106 (2015)
7. Wisniowska, B., Mendyk, A., Polak, M., Szlek, J., Polak, S.: Randomforest based assessment of the herg channel inhibition potential for the early drug cardiotoxicity testing. Int. J. Adv. Manuf. Technol. **47**(9), 1023–1031 (2010)
8. Son, J., Jung, I., Park, K., Han, B.: Tracking-by-segmentation with online gradient boosting decision tree. In: IEEE International Conference on Computer Vision, pp. 3056–3064 (2016)
9. Dorfman, R.: A formula for the Gini coefficient. Rev. Econ. Stat. **61**(1), 146–149 (1979)

Improved Performance of Wireless Sensor Network Based on Fuzzy Logic for Clustering Scheme

Trong-The Nguyen[1,3](✉), Jeng-Shyang Pan[1], Shu-Chuan Chu[2], Thi-Kien Dao[1], and Van-Chieu Do[3]

[1] Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, Fuzhou, Fujian, China
jvnthe@gmail.com

[2] College of Science and Engineering, Flinders University, Tonsley, SA 5042, Australia

[3] Department of Information Technology, Hai-Phong Private University, Hai Phong, Vietnam

Abstract. The wireless sensor network (WSN) consists of a large number of sensor nodes collaborative to collect and transmit data to the end user. Since the network's long life is an utmost requirement of WSN. Clustering is one of the most effective ways of prolonging the lifetime of the network. In clustering, a node takes charge of the cluster to coordinate and receive information from the member nodes and transfer it to the sink. With the imbalance of energy dissipation by the sensor node, it may lead to premature failure of the network. Therefore, a robust balanced clustering algorithm can solve this issue in which a worthy candidate will play the cluster head role in each round. This paper proposes an improvement of WSN based on fuzzy logic for clustering. Residual energy, distance from the sink, and density of the nodes in its locality are taken account as the input to feed into fuzzy inference system. Compared results with the other approaches in the literature show the proposed scheme provides the better performance in terms of stability period and protracted lifetime.

Keywords: Energy efficiency · WSN · Cluster formation · Fuzzy logic Cluster head

1 Introduction

In the rapid development of wireless communication, embedded computing and diverse sensor technology, Wireless sensor network (WSN) is emerging very swiftly [1]. WSN consists of enormous of cheap sensors deployed in the interested area collects the data from surrounding and forward it to the sink for further processing by end-users [2]. Applications of WSN have become popular widely in the fields of e.g., defence, structural monitoring, industrial monitoring, environmental monitoring, climatic and weather monitoring, natural disaster, health care etc., [3]. Collaborative work of a large number of sensor nodes in WSN perform the application tasks. Prolong lifetime of the network is an utmost requirement of WSN due to it is resource restricted with regard to

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 104–113, 2019.
https://doi.org/10.1007/978-3-030-04585-2_13

energy, computation and communication [1, 4]. Because sensor nodes are lightweight and tiny devices with low power, their transmission range is restricted to conserve energy. If the transmission data of the node is long distance, the network is generally accomplished by multi-hoping or intermediate nodes [5]. So, the proficient and proper utilisation of sensor nodes power is required for the longer lifespan of the network. In some WSN applications also used the auxiliary resource like solar cells [6], the power source of sensor nodes could be top up, and nonetheless, it is not continuous which would hamper the functioning of the device.

The key addressing issues for prolonging lifetime is network topology and efficient energy consumption. Cluster-based schemes effectively maintain network topology by partitioning the field and forming clusters. For conserving the energy, two constituents emphasised: a number of operations performed by the sensor nodes and the communication method appertained. Most clustering schemes consist of two primary states: electing cluster-head (CH) based on maximum remnant energy level and rotating CHs periodically to balance the power consumption of nodes [7]. However, the CHs are near the base station (BS) would die out earlier, because they are subject to heavier relay traffic load than the CH farther away from the BS that known as hot spots problem.

This paper proposes a new clustering scheme using fuzzy logic to make a further improvement in protracting the network lifespan of the WSN. Residual power and closeness to BS are considered for calculating the chance of selecting the CH, whereas once the nodes choose CH, remnant power and closeness to CH are also used. Both the parameters are crucial in processing clustering, however, the density around the tentative CH is also important in the cluster for leading to more energy dissipation in transmission and receptions of data. The parameters of the node density, remnant power and distance from BS are taken in account while choosing the best candidates for CH role as well as in the selection of tentative CH by the member nodes during cluster formation. For evaluating the performance of the proposed scheme, simulation experiments are carried out and compared with other approaches in the literature.

2 Related Work

Clustering technologies have been figured out for exploration and application in WSN for recently. Some popular of clustering techniques included LEACH (Low-Energy Adaptive Clustering Hierarchy) [7] and its improved versions, CAFL (Fuzzy Logic Based Clustering Algorithm for Wireless Sensor Networks) [8], CHEF (Cluster Head Election mechanism using Fuzzy logic in WSN) [9], and the others are highlighted to review in this section. LEACH is the pioneer protocol for cluster formation of sensor nodes in WSN [7]. It is a distributive protocol that makes decisions locally for the selection of CH. It is randomized in terms of rotation of the CH role to distribute the load evenly. This protocol also performs data compression at the CH level so as to minimize the amount of data directed to BS. An improved LEACH that altered the threshold for best CH candidate selection and modified the TDMA schedule for better transmission mechanism [10]. First implemented Fuzzy based clustering approach was an improvement of LEACH that the network lifetime was efficiently increased by using three input constituents (node centrality, the degree of the node and its energy) in Fuzzy Inference System (FIS) [11]. Twenty-seven fuzzy if-then rules are used for selecting CH by BS applying the rules. However, this approach was centralized, therefore it is not suitable for scalable networks.

CAFL [8], another approach ensured load balancing by forming clusters using fuzzy logic. There are three input parameters to fuzzifier: node density, distance from BS and the remaining energy of the node, and output parameters are chance and size. Experimental results portrayed better performance of this algorithm than its comparatives. CAFL is a distributed algorithm which assigned the predefined probability to sensor nodes as per their farness from BS. It used fuzzy logic to select optimal candidates for CH role.

CHEF [9] used two input variables for FIS: node's remnant energy and local distance to evaluate the fuzzified inputs and calculation of node's chance elected as coordinator of the cluster [9]. Nine fuzzy rules of IF-THEN were used to select a tentative CH by calculating the competitive radius to compete for CH candidature. However, this method did not contemplate energy dissipation due to huge intra-communication which deteriorates the performance of the protocol. HEEDML-FL is an enhancement over HEED [12] with fuzzy logic incorporation included remnant energy and node density [13].

This paper emphasizes on adapted fuzzy logic to form clusters with parameters like density, energy and farness from BS and substantially improves the lifetime of WSN. The metric based on remnant energy, dead nodes, first dead node, half dead node and last dead node is used as the criteria for comparison to evaluate the proposed scheme that is presented in Sect. 4.

3 Proposed Approach

3.1 Parameters Setting

Assumption WSN environment has been made in the implementation of the proposed approach, as follows:

1. The sensor nodes are deployed randomly.
2. The sensor nodes are homogeneous with battery level at parity.
3. The base station, as well as sensor nodes, are immobile.
4. The power supply to sensor nodes is irreplaceable and non-rechargeable whereas Base station has a continuous power supply.
5. The separation distance is computed by Link quality indicator (LQI).
6. The communication between any two devices is symmetric.
7. The sensor nodes will be presumed dead only if its power supply gets exhausted.
8. The target field is presumed to be the area of 100×100 units.

The symbol for assumed parameters is used in setting simulation as shown in Table 1.

Table 1. Description symbol for assumed parameters

Description	Symbol	Values
Total sensor nodes in the field of a network area	NS	100
Amplifier energy for free space	ε_{fs}	10 pJ/bit/m^2
Amplifier energy for multipath	ε_{mp}	0.0013 pJ/bit/m^4
Battery level before deployment	Eo	0.5 J/1.0 J
Data packet size	M	4000 bits
Energy for data fusion	E_{DA}	5 nJ/bit/report
Energy for electronic circuitry	E_{elec}	50 nJ/bit

3.2 Radio Energy Model

The radio energy model was defined by the amount of energy dissipated for transmission and reception of m bits of the message as represented as E_{Tx} and E_{Rx} respectively.

$$E_{Tx}(m,d) = \begin{cases} mE_{elec} + m\varepsilon_{fs} \times d^2, & d < d_0 \\ mE_{elec} + m\varepsilon_{mp} \times d^4, & d \geq d_0 \end{cases} \tag{1}$$

where d is a distance of Transmitter and Receiver; ε_{fs} and ε_{mp} are the energy loss of free space and multi-path model respectively; d_0 is a threshold as definded as $d_0 = \sqrt{\varepsilon_{fs}/\varepsilon_{mp}}$. For receiving m-bits data, the energy required is calculated as follows.

$$E_{Rx}(x) = E_{Rx-elec}(m) = s.E_{elec} \tag{2}$$

where E_{elec} is the consuming energy for transmitter and receiver of the circuitry per bit. The total energy exhausted for a CH (cluster head) in each round is calculated as follows.

$$E_{CH} = nm(E_{elec} + \varepsilon_{fs}d_{BS} + E_{DA} \tag{3}$$

where d_{BS} is the distance between CH and BS; n is the count of member nodes. Similarly, the amount of energy dissipated by a cluster member is computed as follows:

$$E_{CM} = m\left(E_{elec} + \varepsilon_{fs}d_{CM}\right) \tag{4}$$

where d_{CM} is the distance to its CH.

3.3 Selecting CH Based on Incorporating Fuzzy Logic

Once the sensor nodes scattered in the area of interest, the proposed scheme comes into play with having two phases: CH selection and Cluster formation. Fuzzy logic is incorporated for selecting CH candidate. The decision-making behaviour is efficiently handled by calculating the rank of each node. The proposed method works based on rounds as iterations. Three input variables; density, remnant energy and farness from

BS are applied to FIS. The node density is also taken into account while making any decision. The overhead of the CH is also considered while electing it as CH. The decision is made on the basis of the rank. Figure 1 shows the clustering inference system. The Mamdani inference system is adopted which is most commonly used because of its characteristics and simplicity. FIS endeavours to imitate the human inference system in making a conclusion from the given set of constraints in the knowledge base. The procedure of defuzzification maps the fuzzy set obtained from the inference engine into crisp value for drawing the conclusion. Table 2 lists the inputs and linguistic variables of the fuzzy logic for ranking CH selection.

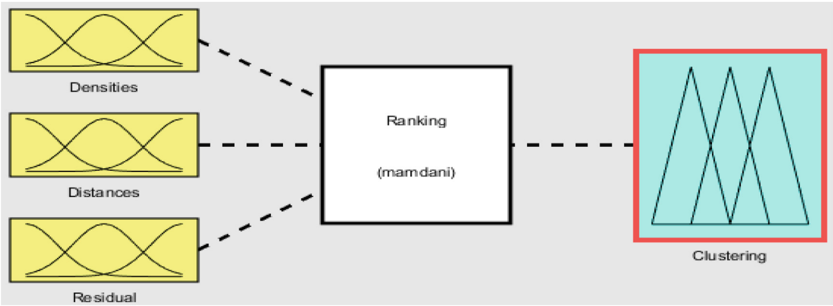


Fig. 1. The clustering fuzzy inference system (FIS)

The membership function (MF) are framed by satisfying a condition that it should range from 0 to 1 that is crisp input values. The MF of Triangular and Trapezoidal are used for intermediate and boundary variables respectively since it is simpler with faster computation [14]. The values obtained after fuzzification are provided to rule base to test for IF-THEN conditions. There are twenty-seven fuzzy rules for node ranking as depicted in Table 2.

The values obtained after fuzzification are provided to rule base to test for IF-THEN conditions. Because of the permutation of the three input variables with each variable having three linguistic parameters, so the total fuzzy rules can be twenty-seven (3^3 = 27 linguistics). Table 3 illustrates a brief listing of twenty-seven fuzzy rules for node ranking. Output linguistic variables for ranking are able being Very Strong, Strong, Rather Strong, Medium Strong, Medium, Medium Weak, Rather Weak, Weak, and Very Weak.

The procedure of defuzzification maps the fuzzy set obtained from the inference engine into crisp value for drawing a conclusion. Table 3 indicates the fuzzy variables used are for crisp output. Defuzzifier changes the obtained input from the inference engine into the crisp set using triangular and trapezoidal MF. Once all the nodes are ranked, the node would broadcast the rank within transmitting range as per iterations. The node's ranks are sorted, if a node's rank is highest, it will proclaim its candidature for CH by broadcasting a packet (HEADING) which contains its density and residual energy otherwise it will wait to join the optimal cluster. Thus, the candidates with higher rank will be elected as CH. Algorithm 1 shows the steps of selecting the cluster head (CH) based on Incorporating Fuzzy logic.

Table 2. Variables of the fuzzy logic for ranking CH selection

Input variables	Linguistics		
Node densities	High	Medium	Low
Closeness to base station	Far	Medium	Near
Residual energies	High	Medium	Low

Table 3. Fuzzy rules for Nodes ranking

Rules	Residual energy	Closeness to base station	Node densities	Ranking radius
1	*High*	*Far*	*High*	*Strong*
2	*High*	*Medium*	*High*	*Very strong*
..	.	.	.	.
..	.	.	.	.
27	*Low*	*Medium*	*High*	*Rather weak*

Algorithm 1. Selection CH candidate

```
Begin
SN ← Overall SN in the field
i ← Unique Identity
SN(i).Energy← current power level
SN(i).Type ← member
SN(i).Rank ← 0
List_CH ← 0
Count_CH ← 0
SN(i).Density ← Total SN inside Transmitter Range (Rc)
SN(i).DBS ← Distance from BS to SN(i)
Compute rank of each Node (i).
Broadcast rank to proclaim CH candidature
while (Count_CH <=p% ) do
if SN(i).Rank > Received_SN(j).Rank then
SN(i).Type ← CH
Count_CH ++
Add SN(i) to List_CH
Broadcast (HEADING) packet
end if
end while
End
```

3.4 Cluster Formation

The cluster is formed after the CH selection phase, the member nodes have to make a decision to join one of candidate CHs. The joining decision of nodes would be made by calculating the chance to which CHs by fuzzy inference with three inputs included Density, Closeness to CH, and CH residual energy (Table 4).

Table 4. Variables of the fuzzy logic for measuring CH chance

Input variables	Linguistics		
CH densities	High	Medium	Low
Closeness to CH	Far	Medium	Near
CH residual energies	High	Medium	Low

A cluster could be formed if CH is the highest chance, the node will send a (JOINING) packet to the CH for joining requestion. The CH will accept the request by sending an acknowledgement packet to the node for confirmation. In this way, all clusters are successfully formed. Afterwards, CH will send TDMA slot to all member nodes to collect data evading collision. After the collection of information, CH fuses the data for minimizing the communication cost. The fused data will transfer to BS for further processing by the end user. Algorithm 2 illustrates the steps of forming the cluster.

Algorithm 2: Clustering Formation

Begin:

1: *REC_CH_LIST ← All CH whose packet (CL_HEAD) is received.*
2: *i ← Distinct ID of sensor nodes*
3: *Compute chance of each CH_Node in REC_CH_LIST*
4: *OPT_CH ← 0*
5: ***while** (End of REC_CH_List) **do***
6: ***if** CH_Node(j).Chance > OPT_CH **then***
7: *OPT_CH ← j*
8: ***end if***
9: ***end while***
10: *SN(i).CH←j* //SN with ID as j is selected as CH for SN(i)
11: *SN(i) will send a packet (JOIN_REQ) to SN(j)*
12: *SN(j) will send ACK to SN(i) with TDMA slot.*

End

4 Simulation Results

The simulation experiment is carried out extensively to test the performance of the proposed method according to several categories such as the lifetime, residual energy of the network, and the status of the alive nodes. The obtaining normalised results are compared with the other approaches in literature e.g. LEACH, CHEF and CAFL. Matrix laboratory is used for simulation work as it is easier for Fuzzy logic implementation. The configuration of the simulation and initial parameters setting is listed in Table 1. Figure 2 compares the network's lifetime of the proposed approach with LEACH, CHEF, and CAFL. It depicts the count of nodes still alive in each round. Clearly, the number of rounds of the proposed approach is outstanding, i.e. the proposed approach has a longer time of network lifetime than the other methods.

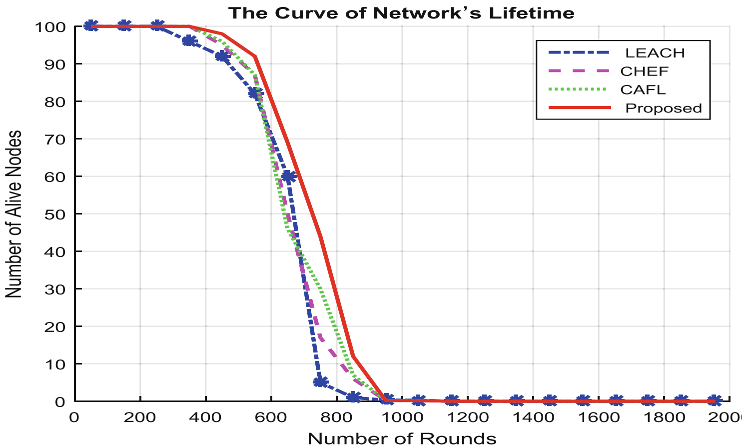


Fig. 2. Comparison of the network lifetime of the proposed scheme with the LEACH, CHEF and CAFL schemes

Figure 3 depicts the residual energy of the proposed approach's network remain energy in comparison with the hierarchical routings as LEACH, CAFL, and CHEF. It shows the total energy of the network available in each round. Noticeably, the proposed method remains the residual energy of the network is higher than the others, i.e., the proposed method consumed power less than the LEACH, CAFL and CHEF methods.

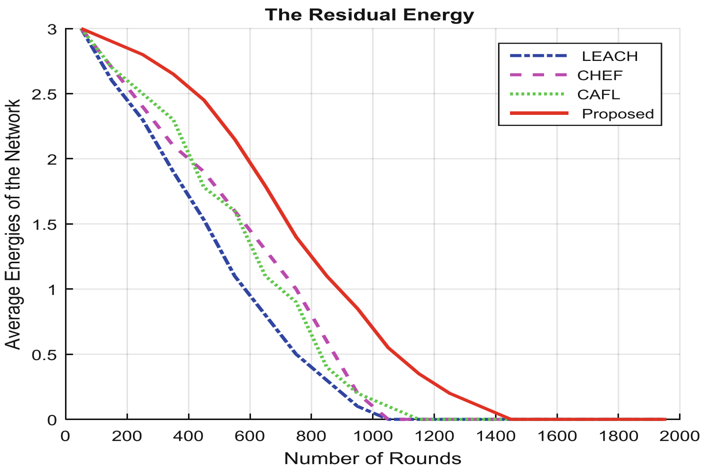


Fig. 3. Comparison of the residual energy of networks of the proposed scheme with the LEACH, CHEF and CAFL schemes

Three metrics for the evaluated lifetime of the network are used by Handy et al. [15]. They are FND (the First node dies), HNA (the Half of the nodes alive) and LND (the Last node dies). The metric of FND is the value that is the first died node. Figure 4 shows the comparison of the proposed approach in rounds that the first node died (FND) half of the nodes alive (HNA) and last node dies (LND) with LEACH, CHEF, and CAFL. Clearly, the proposed scheme provides the outstanding results for the network lifetime.

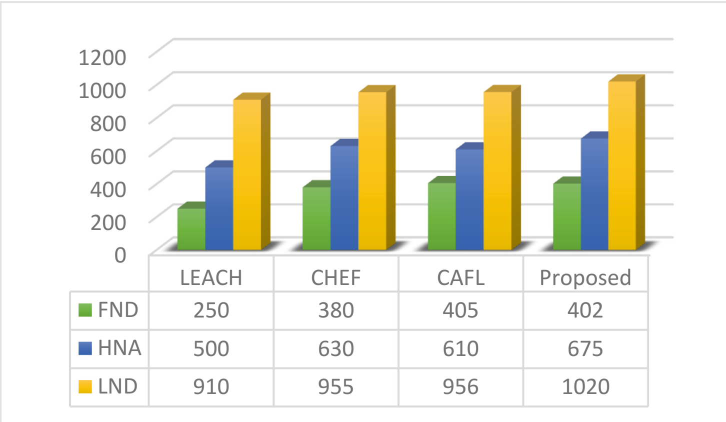


Fig. 4. Comparison of the proposed approach in rounds by the metric e.g. FND, HNA, and LND with LEACH, CHEF, and CAFL

5 Conclusion

In this paper, an improvement for the performance of Wireless sensor network (WSN) clustering based on Fuzzy logic was proposed. The effective factors of the network such as the residual energy, the node density in its locality and the distances from the base station are taken account into a fuzzy inference system. A rank of each node is computed for the candidature of cluster coordinator. Experiments are performed for the proposed scheme to validate its performance. Compared experimental results with the other methods in the literature e.g. LEACH, CHEF, and CAFL schemes shows the proposed scheme provides the better performance in terms of stability period and protracted lifetime.

References

1. García-hernández, C.F., Ibargüengoytia-gonzález, P.H., García-hernández, J., Pérez-díaz, J.A.: Wireless sensor networks and applications: a survey. J. Comput. Sci. **7**, 264–273 (2007)
2. Fan, G., Jin, S., Processing, D.: Coverage problem in wireless sensor network: a survey. J. Netw. **5**, 1033–1040 (2010)
3. Pan, J.-S., Nguyen, T.-T., Dao, T.-K., Pan, T.-S., Chu, S.-C.: Clustering formation in wireless sensor networks: a survey. J. Netw. Intell. **02**, 287–309 (2017)

4. Yick, J., Mukherjee, B., Ghosal, D.: Wireless sensor network survey. Elseiver Comput. Netw. **52**, 2292–2330 (2008)
5. Sabor, N., Sasaki, S., Abo-zahhad, M., Ahmed, S.M.: A comprehensive survey on hierarchical-based routing protocols for mobile wireless sensor networks: review, taxonomy, and future directions. Wirel. Commun. Mob. Comput. **2017**, 23 (2017)
6. Alippi, C., Galperti, C.: An adaptive system for optimal solar energy harvesting in wireless sensor network nodes. IEEE Trans. Circuits Syst. I Regul. Pap. **55**, 1742–1750 (2008)
7. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of 33rd Annual Hawaii International Conference System Science, pp. 3005–3014 (2000)
8. Lee, J.-S., Cheng, W.-L.: Fuzzy-logic-based clustering approach for wireless sensor networks using energy predication. IEEE Sens. J. **12**, 2891–2897 (2012)
9. Kim, J.-M., Park, S.-H., Han, Y.-J., Chung, T.-M.: CHEF: cluster head election mechanism using fuzzy logic in wireless sensor networks. In: 2008 10th International Conference on Advanced Communication Technology, pp. 654–659 (2008)
10. Elshrkawey, M., Elsherif, S.M., Wahed, M.E.: An enhancement approach for reducing the energy consumption in wireless sensor networks. J. King Saud Univ. - Comput. Inf. Sci. **30**, 259–267 (2017)
11. Gupta, I., Riordan, D., Sampalli, S.: Cluster-head election using fuzzy logic for wireless sensor networks. In: Proceedings of 3rd Annual Communication Networks and Services Research Conference (2005)
12. Younis, O., Fahmy, S.: HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. IEEE Trans. Mob. Comput. **3**, 366–379 (2004)
13. Singh, S., Chand, S., Kumar, B.: Energy efficient clustering protocol using fuzzy logic for heterogeneous WSNs. Wirel. Pers. Commun. **86**, 451–475 (2016)
14. Nayak, P., Devulapalli, A.: A fuzzy logic-based clustering algorithm for WSN to extend the network lifetime. IEEE Sens. J. **16**, 137–144 (2016)
15. Handy, M.J., Haase, M., Timmermann, D.: Low energy adaptive clustering hierarchy with deterministic cluster-head selection. In: 2002 4th International Workshop on Mobile and Wireless Communications Network, MWCN 2002 (2002)

Research on Power Transformer Selection Based on Blind Number Theory

Renwu Yan(✉), Jiaman Luo, Shuoxun Gao, and Chuan Lin

Fujian Key Laboratory of Automotive Electronics and Electric Drive, College of Information Science and Engineering, FuJian University of Technology, Fuzhou 350118, China
16370379@qq.com, 359312136@qq.com, 295096725@qq.com, 492456598@qq.com

Abstract. The life cycle cost (LCC) of power transformers is the main factor to be considered in the process of investment selection, and there are many uncertain factors in the life cycle cost assessment process of power transformers. In this paper, the blind number theory is introduced into the life cycle cost analysis of power transformers and build a mathematical model. The model includes initial investment cost, operation and maintenance cost, loss cost, environmental cost, failure cost, and decommissioning disposal cost of the power transformer. The life cycle cost of power transformers is analyzed, which provides a reference for power transformer selection. The selection result can not only calculate the blind number expectation value of the life cycle cost of the power transformer, but also obtain the possible distribution interval of different cost and the corresponding credibility information, improve the estimation result and the rationality of power transformer selection.

Keywords: Power transformer · Selection · Life cycle cost
Blind number theory

1 Introduction

Power transformer is one of the most important equipments in power network, which plays an important role in power transmission and voltage adjustment. This paper mainly focuses on the economics of power transformer selection. In the process of power equipment procurement and construction, only the initial investment cost is considered, but the total cost of power equipment in the whole life cycle is ignored. The cost of the latter is often several times of the initial investment cost [1]. Therefore, it is of great theoretical and practical significance to analyze and study the full life cycle cost of power transformer to improve the stable and economic operation of power transformer and provide reference for the decision of power transformer selection.

Experts and scholars have reached a relatively mature stage in the study of power transformer selection based on full life cycle cost. The literature [2] established a more detailed transformer LCC model and comprehensively analyzed various cost models. The literature [3] proposed the concept of comprehensive evaluation indicators, and selected transformers through comparison of comprehensive evaluation indicators of

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 114–122, 2019.
https://doi.org/10.1007/978-3-030-04585-2_14

various types of transformers. The literature [4] proposes the application method of the full life cycle cost theory in the selection of main transformer, and its feasibility is proved by an example analysis.

Based on the blind number theory and the time value of funds, this paper establishes a mathematical model of the life cycle cost of power transformers. This paper analyzes and solves the uncertainty information in the whole life cycle cost estimation of power transformer effectively, and obtains the possible distribution interval of the whole life cycle cost of power transformer and the reliability information of the corresponding interval. The application of full-life cycle cost theory in power transformer selection is presented. And the feasibility is proved by example analysis.

2 Power Transformer Cost System from the Perspective of Life Cycle

Life cycle cost of power transformers refers to the total cost incurred in the initial purchase cost, installation and commissioning cost of power transformers, as well as the operation and maintenance cost, loss cost, environmental cost, failure cost, decommissioning and disposal cost of power transformers.

2.1 Initial Investment Cost of Power Transformer

The initial investment cost refers to the total cost incurred by the equipment from the procurement to the first operational phase, which is mainly caused by equipment purchase costs, equipment installation and commissioning costs, and other expenses.

$$\mathrm{CI} = \left(C_{gz} + C_{az} + C_{qt}\right) \cdot \frac{i(1+i)^n}{(1+i)^n - 1} \tag{1}$$

In the formula, C_{gz} represents the purchase cost of the equipment, C_{az} represents the installation and debugging cost of the equipment, and C_{qt} represents other expenses.

2.2 Operating and Maintenance Costs of Power Transformers

Operation and maintenance costs refer to the total cost of operations and maintenance during the period from the initial commissioning to decommissioning. Among them, the operating expenses are mainly composed of equipment and materials costs required for daily operation inspection, operation inspection labor costs. The maintenance cost refers to the maintenance cost due to the failure of the transformer.

$$\mathrm{CO} = \mathrm{CO(x)} \tag{2}$$

In the formula, $\mathrm{CO(x)}$ is the blind value of the annual operation and maintenance cost estimated by the expert.

2.3 Loss Cost of Power Transformer

The loss cost of power transformer refers to the total cost of power transformer in operation due to the loss of power. Power transformer is one of the most energy-consuming equipment in the power grid. The total loss of transformers in China is about 3% to 5% of the total power generation of the power system [5].

$$\mathrm{CW} = \mathrm{CW(x)} \tag{3}$$

In the formula, $\mathrm{CW(x)}$ is the blind value of annual loss cost estimated by experts.

2.4 Environmental Cost of Power Transformers

The environmental cost of power transformers refers to the greenhouse effect gases (CO2, CH4, etc.) and noise pollution caused by power transformers in operation. These environmental impacts are quantified in a certain way and environmental costs are reflected in the full-life cycle cost model of power transformers.

$$\mathrm{CE} = \mathrm{CE(x)} \tag{4}$$

In the formula, $\mathrm{CE(x)}$ is the blind value of the annual environmental cost estimated by the expert.

2.5 Power Transformer Failure Cost

The fault cost of power transformer refers to the loss caused by the influence on normal power supply after the failure of power transformer. It mainly consists of three parts: power supply loss, repair cost and penalty cost.

$$\mathrm{CF} = \mathrm{CF(x)} \tag{5}$$

In the formula, $\mathrm{CF(x)}$ is the blind value of the annual failure cost estimated by the expert.

2.6 Decommissioning and Disposal Cost of Power Transformers

Decommissioning disposal costs refer to the total costs incurred in the process of disposal of equipment when it is decommissioned. It is mainly composed of the scrapped cost of the power transformer and its residual value. The residual value of power transformers is about 20% to 30% of the purchase cost [6].

$$\mathrm{CD} = -\rho \mathrm{CI} \frac{I(1+\mathrm{i})^n}{(1+i)^n - 1} \tag{6}$$

In the formula, CD is the initial investment cost of power transformer; ρ is the recovery factor, which is 30%.

3 Full-Life Cycle Cost Estimation Model of Power Transformer Based on Blind Number Theory

3.1 Basic Theory of Blind Number

Blind number theory was established and developed by liu kaidi and wu heqin on the basis of unascertained mathematical theories, and is a mathematical tool for processing and expressing blind information [7].

Definition of Blind Number
Let G(I) be an interval-type rational gray number set, $x_i \in G(I), \alpha \in [0,1]$, $\mathrm{i} = 1, 2, \cdots, \mathrm{n}$, $\mathrm{f}(x)$ is the gray function defined on $\mathrm{G}(I)$ [8–10], and has:

$$\mathrm{f(x)} = \begin{cases} \alpha_i, x = x_i(i = 1, 2, \cdots, n) \\ 0, other \end{cases} \tag{7}$$

α_i is the credibility of $\mathrm{f(x)}'\mathrm{s}\ x_i$ value, and $\sum_{i=0}^{n} \alpha_i$ is the total credibility of $\mathrm{f}(x)$ [15].

When $\mathrm{i} \neq \mathrm{j}$, $\mathrm{x}_i \neq x_j$, and $\sum_{i=0}^{n} \alpha_i = \alpha \leq 1$, then $\mathrm{f}(x)$ is called a blind number, and n is the order of $\mathrm{f}(x)$.

Blind Number Operation
Let the blind numbers A and B be:

$$\mathrm{A} = \mathrm{f(x)} = \begin{cases} \alpha_i, x = x_i(i = 1, 2, \cdots, n) \\ 0, other \end{cases} \tag{8}$$

$$\mathrm{B} = \mathrm{g(y)} = \begin{cases} \beta_j, y = y_j(j = 1, 2, \cdots, n) \\ 0, other \end{cases} \tag{9}$$

Then the blind number four operations of A and B are realized by the possible value edge * matrix and the credibility edge product matrix.

Mean Value of Blind Number
Assuming that a and b are both real numbers and a $\leq$ b, then the process of (a + b)/2 is recorded as $\theta(\mathrm{a}, \mathrm{b})$. The mean value of the blind number A and B is calculated as:

$$\mathrm{E(A)} = \mathrm{Ef(x)} = \begin{cases} \alpha_i, x = \sum_{i=1}^{m} \alpha_i \cdot \theta(x_i, x_{i+1}) \\ 0, other \end{cases} \tag{10}$$

$$\mathrm{E(B)} = \mathrm{Ef(y)} = \begin{cases} \beta_j, y = \sum_{j=1}^{m} \beta_j \cdot \theta(y_j, y_{j+1}) \\ 0, other \end{cases} \tag{11}$$

Blind Number BM Model
Suppose A and B are blind numbers and q is a real number determined in the problem.

When the evaluation index is the larger the better, the blind BM model is:

$$\mathrm{P}(\mathrm{A}-\mathrm{B}\geq \mathrm{q}) = \sum\nolimits_{x_i - y_{j\geq q}} f(x_i)\mathrm{g}(y_i) \tag{12}$$

3.2 Mathematical Model

According the formulas (1) to (6), the blind number expression of the life cycle cost of the power transformer:

$$\mathrm{LCC} = \mathrm{CI}(1-\rho)\frac{i(1+i)^n}{(1+i)^n - 1} + CO(x) + CW(x) + CE(x) + CF(x) \tag{13}$$

4 Case Analysis

This paper takes the power transformers of two different manufacturers as an example. Firstly check the basic data (Table 1) and historical related data of the power transformers of these two schemes, and invite the experts in the industry to make a qualitative estimate of the costs of the two schemes [11].

Make the following assumptions:

(1) The expected life of the transformer is equal to 30 years, and the annual interest rate is 8% and load rate is 0.65;
(2) The residual value is considered at 30% of the transformer purchase cost;

Table 1. 220 kV main transformer research basic data

Transformer	Plan 1	Plan 2
Manufacturer types	Domestic	Foreign
Initial investment cost (million)	4.85	6.25
No-load annual loss (kW)	35.18	27.23
Annual load loss (kW)	233.58	218.62
The total capacity (MVA)	120	120

(1) Initial investment cost of power transformer:

According to the information, the initial investment cost of Option 1 is 4.85 million yuan and Option 2 is 6.25 million yuan. Therefore, after the conversion to the same annual value, the initial investment of the two programs is:

$$CI_1 = 4.85 \times \frac{8\%(1+8\%)^{30}}{(1+8\%)^{30}-1} = 0.4308 \; million$$

$$CI_2 = 6.25 \times \frac{8\%(1+8\%)^{30}}{(1+8\%)^{30}-1} = 0.5552 \; million$$

(2) Power transformer decommissioning disposal cost:

According to the formula, the disposal cost of the power transformer is:

$$CI_1 = -(4.85 \times 30\%) \times \frac{8\%(1+8\%)^{30}}{(1+8\%)^{30}-1} = -0.1292 \; million$$

$$CI_2 = -(6.25 \times 30\%) \times \frac{8\%(1+8\%)^{30}}{(1+8\%)^{30}-1} = -0.1666 \; million$$

Taking the loss cost of power transformers as an example, three experts made a qualitative estimate of the loss costs of the power transformers of the two schemes. The specific results are as follows:

(1) Expert A estimates that the annual loss costs of power transformers in the first scheme and the second scheme are 1.35 to 1.45 million and 1.20 to 1.30 million respectively, which is estimated to be 75% certain.
(2) Expert B estimates that the annual loss costs of power transformers in the first scheme and the second scheme are 1.45 to 1.55 million and 1.30 to 1.40 million respectively, which is estimated to be 90% certain.
(3) Expert C estimates that the annual loss costs of power transformers in the first scheme and the second scheme are 1.55 to 1.65 million and 1.40 to 1.50 million respectively, which is estimated to be 80% certain.

First, university professors, power company maintenance personnel, and transformer equipment manufacturers are invited to estimate the various costs incurred by the two-factor transformers throughout their life cycle. Combine the academic authority, professional title of the three experts, familiarity with the life cycle cost of the transformer, etc. [11–13]. It is concluded that the weights of the three experts in the cost estimation process are: 0.25, 0.55, 0.20, as shown in Table 2:

Table 2. Expert weight value

Experts	Weight of academic authority	Title weight	Familiarity weight	The total weight value
A	Doctor 0.10	University professor 0.07	0.08	0.25
B	Master 0.05	Maintenance staff 0.10	0.40	0.55
C	Undergraduate 0.02	Factory personnel 0.03	0.15	0.20

Secondly, considering the degree of estimation of each expert, we can get the comprehensive credibility α_1 of expert A, the comprehensive credibility α_2 of expert B, and the comprehensive credibility α_3 of expert C, as follows:

$$\alpha_1 = \frac{\frac{75\% \times 0.25}{75\% + 90\% + 80\%}}{\frac{75\% \times 0.25}{75\% + 90\% + 80\%} + \frac{90\% \times 0.55}{75\% + 90\% + 80\%} + \frac{80\% \times 0.20}{75\% + 90\% + 80\%}} = 0.223$$

$$\alpha_2 = \frac{\frac{90\% \times 0.55}{75\% + 90\% + 80\%}}{\frac{75\% \times 0.25}{75\% + 90\% + 80\%} + \frac{90\% \times 0.55}{75\% + 90\% + 80\%} + \frac{80\% \times 0.20}{75\% + 90\% + 80\%}} = 0.588$$

$$\alpha_3 = \frac{\frac{80\% \times 0.20}{75\% + 90\% + 80\%}}{\frac{75\% \times 0.25}{75\% + 90\% + 80\%} + \frac{90\% \times 0.55}{75\% + 90\% + 80\%} + \frac{80\% \times 0.20}{75\% + 90\% + 80\%}} = 0.189$$

Thus, the blind number expression of the loss cost of the power transformers of two different manufacturers can be obtained:

$$CW_1 = \begin{cases} 0.223, [135, 145] \\ 0.588, [145, 155] \\ 0.189, [155, 165] \end{cases} CW_2 = \begin{cases} 0.223, [120, 130] \\ 0.588, [130, 140] \\ 0.189, [140, 150] \end{cases}$$

In the same way, the possible value range of power transformer other cost and its corresponding comprehensive credibility can be obtained.

The blind number expression for the power transformer Operation and maintenance cost is:

$$CO_1 = \begin{cases} 0.218, [25, 30] \\ 0.622, [30, 35] \\ 0.160, [35, 40] \end{cases} CO_2 = \begin{cases} 0.218, [30, 35] \\ 0.622, [35, 40] \\ 0.160, [40, 45] \end{cases}$$

The blind number expression for the environmental cost of a power transformer is:

$$CE_1 = \begin{cases} 0.158, [4, 6] \\ 0.636, [6, 8] \\ 0.206, [8, 10] \end{cases} CE_2 = \begin{cases} 0.158, [2, 4] \\ 0.636, [4, 6] \\ 0.206, [6, 8] \end{cases}$$

The blind number expression for the power transformer fault cost is:

$$CF_1 = \begin{cases} 0.216, [20, 25] \\ 0.486, [25, 30] \\ 0.298, [30, 35] \end{cases} CF_2 = \begin{cases} 0.216, [5, 10] \\ 0.486, [10, 15] \\ 0.298, [15, 20] \end{cases}$$

Substituting these expressions into the blind number expression (13) of the power transformer. And using BLCC5.3 software for calculation, the result expression after finishing is:

$$CW_{LCC1} = \begin{cases} 0.206, [193.3, 205.3] \\ 0.451, [205.3, 220.5] \\ 0.343, [220.5, 244.8] \end{cases} CW_{LCC2} = \begin{cases} 0.206, [166.4, 178.3] \\ 0.451, [178.3, 193.5] \\ 0.343, [193.5, 217.8] \end{cases}$$

Substituting into Eq. (10), (11) can find the mean of the blind number:

$$C_{LCC1} = \begin{cases} 1, x = 216.88 \\ 0, other \end{cases} C_{LCC2} = \begin{cases} 1, x = 189.88 \\ 0, other \end{cases}$$

Its blind number BM model is:

$$\mathrm{P(A - B \geq 0) = 85.1\%}$$

According to the above calculation results, the total life cycle cost of scheme 1 with a reliability of 0.206 is 1.933–2.053 million yuan. The whole life cycle cost with a reliability of 0.451 is 2.053–2.205 million yuan. The full life cycle cost with a credibility of 0.343 is 2.205–2.448 million yuan. The total life cycle cost of scheme 2 with a reliability of 0.206 is 1.664–1.783 million yuan. The whole life cycle cost with a reliability of 0.451 is 1.783–1.935 million yuan. The full life cycle cost with a credibility of 0.343 is 1.935–2.178 million yuan.

The blind number average of the power transformer LCC of scheme 1 is 2.169 million yuan, and the blind number average of the power transformer LCC of scheme 2 is 1.189 million yuan. From the BM model analysis, the probability that the cost of the scheme 1 is higher than that of the scheme 2 is 85.1%. Although Option 2 has a higher initial investment cost than Option 1, Scheme 2 is clearly superior to Option 1 for all costs after the start of operation.

5 Conclusions

This paper analyzes the life cycle cost mainly considered in the process of power transformer selection, and introduces the blind number theory into the LCC estimation model of power transformer. It fully considers the influence of blind information on the cost estimation result, and can not only evaluate the power transformer selection. The type problem can also obtain the reliability of the possible distribution interval of the power transformer LCC and its corresponding interval and the mean value of its blind number, which is more reasonable than the traditional deterministic estimation algorithm.

References

1. Huijuan, Z., Qinsheng, Z., Ying, W., et al.: Life cycle cost estimation of power switching equipment based on blind number. High-Voltage Electr. Appl. **49**(11), 12–16 (2013)
2. Yuqin, X., Zheng, R., Xiangling, Z., et al.: Lifecycle cost modeling and comprehensive sensitivity analysis of power transformer. J. North China Electr. Power Univ. **41**(6), 80–87 (2014)
3. Huang, L., Ping, W.: Study on power transformer selection method based on comprehensive evaluation index. Zhejiang Electric Power **05**(03), 11–16 (2014)
4. Song, W., Yao, J., Wang, J., et al.: Application of full-life cycle cost theory in main transformer selection. Power Syst. Autom. **24**(6), 111–116 (2012)
5. Jiang, X., Wu, W., Chen, X.: Application of interval analysis method in full-life cycle cost model of power transformers. **31**(9), 50–53 (2011)
6. Jiang, Y., Ma, J.: Life cycle cost analysis of transformers. Transformers **43**(12), 30–34 (2006)
7. Liu, K.: Mathematical processing and application of uncertain information. Science Press, Beijing (1999)
8. Liu, B., Peng, J.: Uncertainty theory course. Tsinghua University Press, Beijing (2005)
9. Liu, B.D.: Uncertainty Theory: An Introduction to its Axiomatic Foundation. Springer, Berlin (2004)
10. Liu, B.D.: Theory and Practice of Uncertain Programming. Physical Verlag, Heidelberg (2002)
11. Wang, G.: Unascertained information and its mathematical treatment. J. Harbin Inst. Arch. Eng. **23**(4), 1–9 (1990)
12. Lu, S., Li, X.: Estimation of total life cycle cost of substation project based on blind number theory. Power Grid Technol. **34**(3), 141–145 (2010)
13. Lu, H.: Preliminary risk study on industrial real estate development based on blind number theory. School of Management, Harbin University of Technology, Harbin (2015)

Dynamic Traffic Information Estimation Based on Floating Car Data

Xiang Xu and Rong Hu(✉)

Fujian Province Key Laboratory of Automotive Electronics and Electric Drive, Fujian University of Technology, Fuzhou 350108, China
896410521@qq.com

Abstract. The road traffic information changed by authority frequently which effect the public travel due to they can't obtain this information timely. Fortunately the floating car data can be used to analyze and estimate such traffic information timely. This work first proposed an advanced map matching method combined both local and global road map information. Second we presented an automatic algorithm based on the substance of road intersection to estimate the road traffic control information in real time. We take the floating car data of Fuzhou city to verify the efficient of our method. Experiments result indicates that the traffic information on intersection can be estimated accurately.

Keywords: Road traffic control information · Floating car data
Map matching · Dynamic traffic information

1 Introductions

Floating cars can collect traffic information without roadside equipment. The GPS sensors equipped in car can send the information of position and speed of car timely in huge volumes of spatio-temporal data in the form of trajectories. Such data provides the opportunity of discovering usable knowledge about movement behavior, which fosters a large range of novel applications and services [1, 15]. Considerable concerns focus on building intelligent transportation system (ITS) whose main purpose is on efficiently managing the road network. Aim to improve the performance of transportation so as to extend the durability of vehicle which can be reduced the fuel consumption and travel time. For instance, many intelligent transportation system (ITS) and services have been designed and developed by several enterprises [2–5]. For the development of ITS, studying how to collect the traffic information efficiently is a really important topic. There are three kinds of popular position collection approaches, which are the vehicle detector, the reports of global positioning system (GPS)-equipped probe cars and cellular floating vehicle data. Vehicle data is a sensor based on the techniques of active infrared/laser, magnetic, radar or video to regularly detect vehicles on a road for analysis of time mean speed and traffic flow [6, 7] GPS-equipped probe cars can periodically report their location information to sever for computing the space mean speed [8]. Floating car data (FCD), which is obtained by tracking the network signals of the mobile station in the car, can be analyzed to estimate the traffic information (e.g., traffic flow, vehicle speed and traffic density).

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 123–130, 2019.
https://doi.org/10.1007/978-3-030-04585-2_15

Some studies discussed traffic collection and estimation methods based on the reports of GPS-equipped probe cars, which were compared to the traffic information from vehicle data for the evaluation of traffic estimation accuracies. For instance, Cheu [9] showed that using active probe vehicles to estimate space mean speed can obtain lower error rate. Herrera [10] showed that GPS-equipped probe vehicles periodically sent their speeds and locations in practical environments which can be analyzed for traffic information estimation.

Cellar floating vehicle data can be obtained by tracking mobile system signals (call arrival. Handover, double handover, normal location update (NLU) and period location update (PLU)), ant it can be used for the analysis of traffic status [11, 12]. This study proposes analytic models and traffic information estimation methods based on cellar floating vehicle data to obtain a low-cost solution for ITS. The proposed analytic models are used to analyze the relationship between traffic information (e.g., traffic density and vehicle speed) and the amount of cellular network signals. Then, traffic information can be estimated based on the proposed analytic models and cellular floating vehicle. Furthermore, a vehicle speed forecasting method based on a back-propagation neural network algorithm is presented to analyze the estimated vehicle speeds from cellular floating vehicle data and to predict the future vehicle speed for road users. Some literatures [13, 14] explored detecting methods on the change of road network by analyzing float car trajectory data. What the area of these researches focus on is about detection or prediction on road traffic condition.

However some other important road traffic condition is not involved in those researches. Road passing limit rule, such as turn-limits, left-turn limit, right-turn limit, one-way passing road and height limit passing, is changed frequently by traffic administrant department due to traffic flow control or some other reasons. This information is the important component of the traffic information, no less important than the congesting information to a driver. This work explores an automatic algorithm to identify this kind of road traffic information in real time and distribute to the public just like the congestion information.

The remainder of this work is organized as follows. Section 2 discusses the method of map matching method. Section 3 presents an automatic algorithm to estimate the instant traffic information. Experimental evaluation and some analysis are made in Sect. 4. In Sect. 5 conclusions are given.

2 Map Matching

Map matching is a process to convert a sequence of raw latitude/longitude coordinates to a sequence of road segments. The first step is to decide which road a vehicle was/is on using floating car data to estimation traffic information. It is important for assessing traffic flow, guiding the vehicle's navigation, predicting where the vehicle is going, and so forth. However, map matching is not an easy problem, given parallel roads, overpasses, and spurs [15]. There are two approaches to classify map-matching methods, based on the additional information used, or the range of sampling points considered in a trajectory. We introduce advanced algorithms embrace local and global information (geometric and topological) to deal with map of a trajectory.

We first define the concepts used in this paper and then give our grid-base map matching algorithm.

Define 1. GPS trajectory: A GPS trajectory T is a point sequence $p_1 \rightarrow p_2 \rightarrow \ldots p_n$ of the GPS points linked with the time stamps. p_i is a triple $(p_i.lat, p_t.long, p_i.t)$ which are its latitude, longitude and timestamp respectively.

Define 2. Road network: A road network is a directed graph G(V, E) where V is a set of vertices representing the intersections and terminal points of the road segments. E is a set of edges representing the road segments. Each road segment e is a directed edge that is associated with an id e.eid.

Define 3. Road Grid: Divide the real road graph into grid $G_i(lo_i, la_i)$ $i = 1, 2, \ldots n$, while is refer longitude and latitude respectively, the size of Grid is about 5 meters.

First dividing the real road map into grid, we scan the road map and mark the grid with segment or none.

Given a GPS trajectory $p_1 \rightarrow p_2 \rightarrow \ldots p_n$, we retrieve a set of candidate road segment of candidate road segments (CRS) for each sampling point that fallen in a grid G_p. The set CRS_i contains all candidate road segments that fallen within grid G_p as shown in Fig. 1.

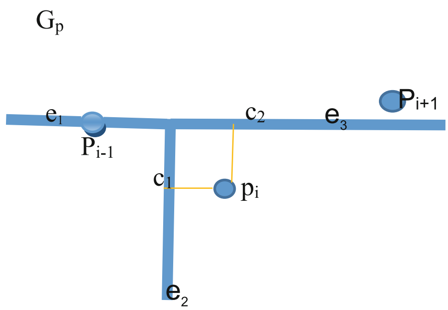


Fig. 1. The candidate road segment fallen into grid

In this phase, we perform the Spatial Temporal Analysis and construct a candidate graph $G'(V, E)$. The distribution of the measurement error is assumed to satisfy the Gaussian distribution probability with respect to π is formulated as (1):

$$N(c_i^j) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x_i^j - u)^2}{2\sigma^2}} \tag{1}$$

Where x_i^j is the Euclid distance from candidate c_i^j to sampling point p_t.

In order to consider both the local and global information, so using recent point estimate map matching is not enough. From Fig. 1, the local candidate road segments are within a grid distance to each point in a trajectory. For instance, road segments e_1, e_2 and e_3 are fallen into the grid G_p. If only consider distance to p_i, it might be matched to c_1 which is located in segment e_2. However if we take the next point p_{i+1} into account, then c_2 is more likely to be the true match of p_i. The basic principle of our method is to place the point to the waiting roads, calculate road section projection distance, and consider driving direction. Introducing matching function as formula (2):

$$S = C_1 \frac{\arctan k - \arctan k_0}{\pi/6} + C_2 \frac{d}{15} \tag{2}$$

C_1 and C_2 represent the weights of angle and distance respectively, $C_1 + C_2 = 1$, this paper $C_1 = 0.5$, $C_2 = 0.5$. k Represents the angle of direction between p_i and p_{i+1}. k_0 Represents the direction angle between p_i and p_{i-1}. d is Euclid distance from candidate to sampling point. Select the smallest S as matching sections, and the projection pints as the vehicle positions after matching. Algorithm process as follows:

Step 1: read information after data preprocessing and data fusion from the received data stream point-by – point and then transform coordinate.

Step 2: Divide the real road map into grid $G_i(lo_i, la_i) \quad i = 1, 2, \ldots n$, while lo_i, la_i refer longitude and latitude respectively, the size of Grid is about 5 m. Scan the road map and mark the grid with segment or none. According to mark of grid, deciding which road segments are the candidate segments for sampling points.

Step 3: Given a GPS trajectory T, for sampling points p_i, we retrieve a set of candidate road segment of candidate road segments (CRS) for each sampling points that fallen in a grid. Mapping sampling point p_i to the candidate road segments according formula (2). Then turn to the next sampling points until all trajectory points are mapped into road segments. The mapping process is over.

3 Traffic Information Estimation

Nearby the road intersection, the traffic condition information includes some driving constraints, such as left-turn limit, right-turn limit, direction limit, no back-turn limit, Speed-limit, and height limit passing. It is changed frequently by traffic administrant department due to traffic flow control or some other reasons. This information is the important component of the traffic information, no less important than the congesting information to a driver. This information can't be reach to the public in time. Fortunately GPS device equipped in floating car send the position information of moving car including latitude and longitude, speed, road segment traveling time and direction. Take the floating cat trajectory data of Fuzhou city for example, the structure as shown in Fig. 1. The data attributes include "MIMD" is the identifier of sensors, the "AREAID" indicate the area the car belong to, the "LOGITUDE" and "LATITUDE" stands for the position of the moving car, "SPEED" indicate the velocity of moving car, "DIRECT" point out the angle between moving car driving and North, "CAR_TYPE" indicate the type of driving car which may be taxi or bus or trunk and etc., and "STATUS" shows the corresponding meaning of attributes is shown as Fig. 2.

	MDID	AREAID	LONGITUDE	LATITUDE	SPEED	DIRECT	STATUS	GPS_DATE	CAR_TYPE
1	2482866	0	118.109927000	24.481513000	20	5	0	2015/2/3 0:00:55	41
2	3836876	0	119.354637000	26.032568000	65	7	0	2015/2/3 0:00:51	31
3	2482261	0	118.145095000	24.547798000	0	5	0	2015/2/3 0:00:55	41
4	2483309	0	118.127297000	24.495667900	37	4	0	2015/2/3 0:00:55	41
5	2482809	0	118.135043000	24.527552000	50	2	0	2015/2/3 0:00:55	41
6	3068620	0	119.256920000	26.015129000	0	4	0	2015/2/3 0:00:51	31
7	2483904	0	118.110497000	24.506387000	0	2	0	2015/2/3 0:00:55	41
8	2914412	0	118.228920000	24.646150000	67	7	0	2015/2/3 0:00:50	31
9	2483111	0	118.095060000	24.475152000	40	7	1	2015/2/3 0:00:55	41
10	3686713	0	118.716783000	24.797117000	0	0	0	2015/2/2 23:59:50	32

Fig. 2. The structure of data attributes

This paper proposes a sustenance-degree based algorithm to estimate traffic information. The basic idea is consider the sustenance of trajectory of car to road. First we prune the trajectory T, only leave the points that fallen in road intersection. In the end, only leave the start point and the end point of the trajectory as shown in formula (3).

$$T(p_1, p_2, \ldots p_n) \rightarrow T(p_i, p_{i+1}, \ldots p_j) \rightarrow p_i, p_j \tag{3}$$

Given in intersection, two road segments (R_m, R_i) are linked together. So the car in road R_m can drive straight to road R_j no need to go through other road. So the sustenance can be computer using ratio between the number of car on road $R_m(R_i)$ and the number of car on road R_m as formula (4):

$$Sus(R_m, R_i) = \frac{\sum_{n=1}^{k} Link - Sum(R_m - l_n, R_i)}{Sum(R_m)} \tag{4}$$

While $Sum(R_m, R_i)$ and $Sum(R_m)$ are counted by the trajectories when it were map to the road segment combined by the identifier of trajectory vehicle data.

If $Sus(R_m, R_i) \approx 0$ then indicate that going straight along is limited called straight-permit.

If $Sus(R_m, R_i) \approx \pi/2$ then indicate that turning left is permit called left-permit.

If $Sus(R_m, R_i) \approx \pi$ then indicate that turning right is permit called right-permit.

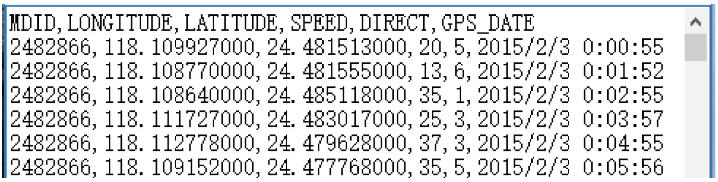
If $Sus(R_m, R_i) \approx 3\pi/2$ then indicate that turning round is permit called U-turn permit.

4 Experiments

To verify the feasibility and effectiveness of road traffic condition recognition algorithm based on floating car, the experiment was made with the software ArcGIS 9.3 and Python 2.7 on a computer with CPU-i5, Windows 7 and 8G RAM. The electronic map used for comparison was the 2006 version of road network data of Fuzhou. In data preprocessing phase, the actual data (as shown in Fig. 3) is transferred into text type as shown in Fig. 3.

The first step is to filter the outlier trajectory data such as the abnormal speed, abnormal time, and abnormal position due to the quality of sensor or the bad environment. According our received data of Fuzhou city, there are about four kinds of noise. The first one is abnormal position which the location of latitude and longitude are same while the speed is nonzero. The second one is that some trajectory points can't be map into the road segments due to some uncertain reason. The third one is the abnormal speed that the speed of car far exceeds the Manxman speed of city permit. The last one is repeat data that the sensor of car sends the same data repeatedly due to the equipment quality or some other uncertain reason. All these types of noise data can be remove by set a reasonable threshold.

The procedure of experiments is depicted as Fig. 4. In order to verify the effect of our method, we select the trajectory data near the intersection road, We special select a road intersection range from longitude [119.26662, 119.26951] to latitude [26.035939, 26.038512] to verify our algorithm using a week about 163394 records of trajectory data. The trajectory data distribute graph is shown as Fig. 5. Figure 4 visualizes the steps of the algorithm. Figure 6a shows the constructed intersection nodes as black stars. The track trajectories distribution of intersection road is shown as Fig. 6. The constituting turn clusters are shown as x and o markers. Figure 6b shows the links between intersections nodes as black lines. The constituting trajectories are shown as dashed lines.



```
MDID,LONGITUDE,LATITUDE,SPEED,DIRECT,GPS_DATE
2482866,118.109927000,24.481513000,20,5,2015/2/3 0:00:55
2482866,118.108770000,24.481555000,13,6,2015/2/3 0:01:52
2482866,118.108640000,24.485118000,35,1,2015/2/3 0:02:55
2482866,118.111727000,24.483017000,25,3,2015/2/3 0:03:57
2482866,118.112778000,24.479628000,37,3,2015/2/3 0:04:55
2482866,118.109152000,24.477768000,35,5,2015/2/3 0:05:56
```

Fig. 3. Text type of floating data

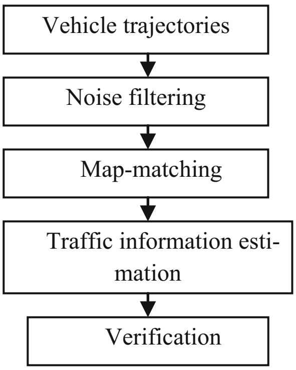


Fig. 4. The procedure of experiments

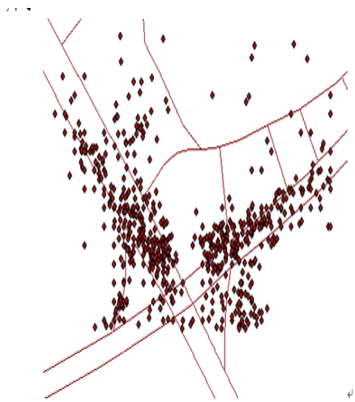
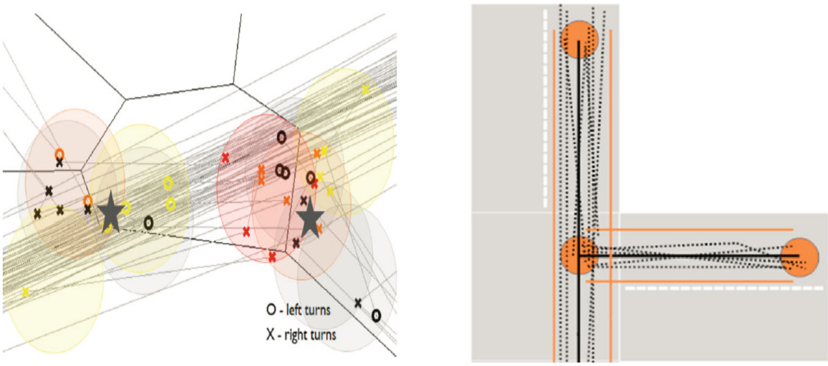


Fig. 5. The trajectory of intersection



(a) (b)

Fig. 6. (a) Intersection nodes and (b) compacting links

In our experiments, the trajectory data contain 30 intersections, and 60 segment roads. Among them there are 3 left-turn limits and 5 right-turn limits 2 U-turn limit that are all recognized by our method. The accuracy rate is 100%.

5 Conclusions

This paper proposes a method to estimate the traffic information through analyzing the floating car data. This kind of information is ignored by experts but very important to intelligent transportation such as intelligent navigation server, travel time prediction, and etc. The contribution of this work mainly focuses on two aspects. First we proposed an advanced map matching method by combining the local and global information of road topology. The nearest distance and the direction of trajectory points are both considered. Second an automatic algorithm for estimating traffic information based on the substance of intersection is presented. Experiment show that our methods are efficiency and the time cost is reasonable. It can be used to detect and estimate the road traffic information timely and efficiency.

Acknowledgments. This work is supported Fujian Provincial Department of Science and Technology, Granted No. 2014H0008, No. 2017J01729, 2017J0106.

References

1. Zheng, Y., Zhou, X.: Computing with Spatial Trajectories. Springer, Heidelberg (2011)
2. Gao, M., Zhu, T., Wan, X., Wang, Q.: Analysis of travel time pattern in urban using taxi GPS data. In: Proceedings of the 2013 IEEE International Conference on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing, 20–23 August 2013, Beijing, China, pp. 512–517. IEEE, Piscataway (2013)

3. Green, D., Gaffney, J., Bennett, P., Feng, Y., Higgins, M., Millner, J.: Vehicle positioning for C-ITS in Australia (Background Document). Project No. AP-R431/13, p. 78. ARRB Group Limited, Victoria, Australia (2013)
4. Ansari, K., Feng, Y.: Design of an integration platform for V2X wireless communications and positioning supporting C-ITS safety applications. J. Glob. Position. Syst. **12**, 38–52 (2013)
5. Schoettle, B., Sivak, M.: A survey of public opinion about connected vehicles in the U.S., the U.K., and Australia. Report No. UMTRI-2014–10, UMTRI, Ann Arbor, MI, USA (2014)
6. Jang, J., Byun, S.: Evaluation of traffic data accuracy using Korea detector testbed. IET Intell. Transport Syst. **5**, 286–293 (2011)
7. Ramezani, A., Moshiri, B., Kian, A.R., Aarabi, B.N., Abdulhai, B.: Distributed maximum likelihood estimation for flow and speed density prediction in distributed traffic detectors with Gaussian mixture model assumption. IET Intell. Transport Syst. **6**, 215–222 (2012)
8. Hunter, T., Herring, R., Abbeel, P., Bayen, A.: Path and travel time inference from GPS probe vehicle data. In: Proceedings of the Neural Information Processing Foundation Conference, Vancouver, BC, Canada, December 2009
9. Cheu, R.L., Xie, C., Lee, D.H.: Probe vehicle population and sample size for arterial speed estimation. Comput.-Aided Civ. Infrastruct. Eng. **17**, 53–60 (2002)
10. Herrera, J.C., Work, D.B., Herring, R., Ban, X.J., Jacoboson, Q., Bayen, A.M.: Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp. Res. Part C: Emerg. Technol. **18**, 568–583 (2010)
11. Chang, M.F., Chen, C.H., Lin, Y.B., Chia, C.Y.: The frequency of CFVD speed report for highway traffic. Wirel. Commun. Mob. Comput. **15**, 879–888 (2015)
12. Yang, J.Y., Chou, L.D., Tung, C.F., Huang, S.M., Wang, T.W.: Average-speed forecast and adjustment via VANETs. IEEE Trans. Veh. Technol. **62**, 4318–4327 (2013)
13. Fang, W., Hu, R., Xu, X., et al.: A novel road network change detection algorithm based on floating car tracking data. Telecommun. Syst. 1–7 (2016)
14. Jiang, X., Liao, L., Zou, F.: Automatic detection algorithm for new roads based on trajectory of floating cars. J. Comput. Appl. **33**(2), 579–582 (2013)
15. Xia, Y., Hu, R.: Fuzzy neural network based energy efficiencies control in the heating energy supply system responding to the changes of user demands. J. Netw. Intell. **2**(2), 186–194 (2017)

Hierarchical Semantic Approximate Multi-keyword Ranked Search over Encrypted Data

Chien-Ming Chen[1], Wenhao Zhang[1], Tsu-Yang Wu[2,3](✉), King-Hang Wang[4], Jimmy Ming-Tai Wu[5], and Jeng-Shyang Pan[2,3]

[1] Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
chienming.taiwan@gmail.com, 327103974@qq.com
[2] Fujian Provincial Key Lab of Big Data Mining and Applications, Fujian University of Technology, Fuzhou 350118, China
[3] National Demonstration Center for Experimental Electronic Information and Electrical Technology Education, Fujian University of Technology, Fuzhou 350118, China
wutsuyang@gmail.com, jspan@cc.kuas.edu.tw
[4] Department of Computer Science and Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
kevinw@cse.ust.hk
[5] College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China
wmt@wmt35.idv.tw

Abstract. With the rapid development of Cloud Computing, how to improve users search experience over encrypted data just like plaintext search has become a challenging task. In this paper, we propose a hierarchical semantic approximate multi-keyword ranked search (HSAMS) scheme over encrypted data. Performance analysis is demonstrated that our proposed scheme is efficient and effective while supporting semantic approximate match.

Keywords: Searchable encryption · Ranked search · Multi-keyword Semantic approximation

1 Introduction

With the rapid development of cloud technologies, some typical cloud service products have also been released with the trend such as Dropbox, Amazon S3, etc. In order to guarantee the confidentiality of outsourced data, a growing number of companies and individual users choose to encrypt data and store the data in the cloud server with ciphertext. However, this approach will lead to a problem that how to search an encrypted data in cloud.

To solve this problem, Song et al. [10] proposed the first searchable encryption (SE) scheme to search a certain keyword in a document, which partially

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 131–138, 2019.
https://doi.org/10.1007/978-3-030-04585-2_16

address the need for encrypted data search. Wang et al. [12,13] first proposed the scheme of ranked keyword search for documents stored in the cloud. Since Wang et al.'s solutions only solve the ranked single keyword search problem, Cao et al. [3] proposed the ranked multi-keyword search (MRSE) scheme for the first time. Later, several literatures [4–8,11,14–21] based on different approaches are designed to solve the search problem in cloud.

In this paper, we take the keywords' semantic relationship between different dimensions into consideration in MRSE-based searchable encryption designs. Based on the keywords' relationship, we design a hierarchical semantic approximate multi-keyword ranked search (HSAMA) scheme over encrypted data. The framework of HSAMA is depicted in Fig. 1. Porter's stemming algorithm and synonym combination are adopted to implement dimension reduction. Meanwhile, a dynamic k-means clustering algorithm is adopted on index vector in order to obtain relatively stable clustering results.

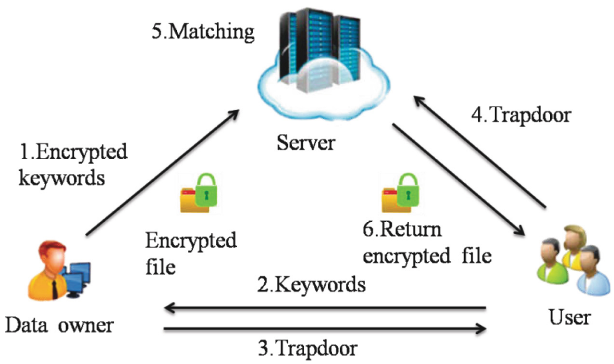


Fig. 1. Framework of HSAMA

2 HSAMA Scheme

Our hierarchical semantic approximate multi-keyword ranked search (HSAMA) scheme consists of following six phases: *Dictionary generation*, *Key generation*, *Index generation*, *Encryption*, *Trapdoor generation*, and *Search*. The used notations are summarized in Table 1.

1. *Dictionary generation.* Data owner executes the following steps to generate dictionary.
   (a) To generate an original keyword dictionary D_1 based on an existing keyword set, where the length of D_1 is m_1.
   (b) Applying the stemming approach [9] to keyword kw in D_1, a stemmed keyword kw' obtained. Thus, a stemmed keyword dictionary D_2 with length m_2 is generated. Meanwhile, a keyword map Map_1 between D_1 and D_2 is generated.

Table 1. Notation

Notations	Meanings
D_1	An original keyword dictionary with length m_1
D_2	A stemmed keyword dictionary with length m_2
D_3	A combined synonym keyword dictionary with length m_3
Γ_1, Γ_2, Λ_1, Λ_2, Ω_1, Ω_2	Random invertible matrices
V_1, V_2, V_3	Random vectors
$\mathfrak{F} = \{F_1, F_2, \ldots, F_n\}$	A collection of n plaintext
I_1	An index vector
I_2	A stemmed index vector
I_3	A synonym index vector
Q_1	A vector with dimension m_1
Q_2	A vector with dimension m_2
Q_3	A vector with dimension m_3
T_{w1}	A trapdoor for exact search
T_{w2}	A trapdoor for derivative words search
T_{w3}	A trapdoor for synonym search

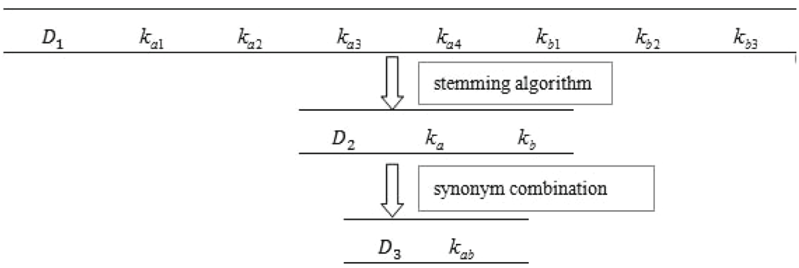


Fig. 2. An example of dictionary generation

(c) Applying the synonym combination approach to kw' in D_2, a combined synonym keyword kw'' obtained. Thus, a combined synonym keyword dictionary D_3 with length m_3 is generated. Meanwhile, a keyword map Map_2 between D_2 and D_3 is generated. Finally, a hierarchical dictionary is generated. An example is depicted in Fig. 2, where k_{a1}, k_{a2}, k_{a3}, k_{a4} (k_{b1}, k_{b2}, k_{b3}) with the same stem and k_a, k_b are synonym.

2. *Key generation.* Data owner generates three private keys $SK_1 = (\Gamma_1, \Gamma_2, V_1)$, $SK_2 = (\Lambda_1, \Lambda_2, V_2)$, and $SK_3 = (\Omega_1, \Omega_2, V_3)$. Note that Γ_1, Γ_2 with size $(m_1 + 2) \times (m_1 + 2)$, Λ_1, Λ_2 with size $(m_2 + 2) \times (m_2 + 2)$, Ω_1, Ω_2 with size $(m_3 + 2) \times (m_3 + 2)$ are random invertible matrices. $V_1 \in \{0, 1\}^{m_1+2}$, $V_2 \in \{0, 1\}^{m_2+2}$, and $V_3 \in \{0, 1\}^{m_3+2}$ are random vectors.

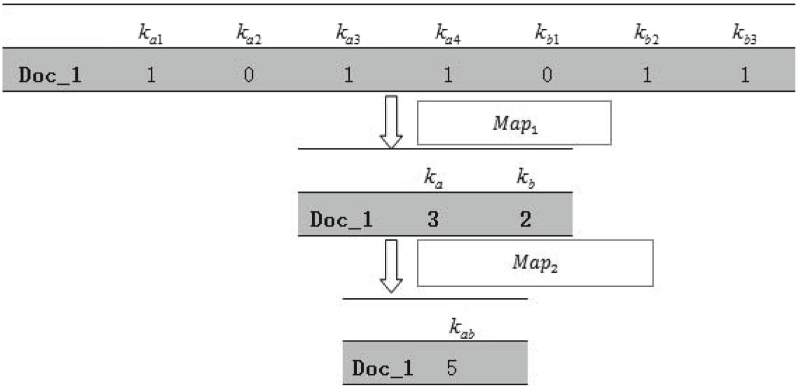


Fig. 3. An example of index generation

3. *Index generation.* Data owner executes the following steps to generate index.
   (a) Based on D_1, the keywords of each plaintext F_i in $\mathfrak{F}$ can be mapped to a vector with size m_1. Thus, we can establish an original index vector I_1 with dimension m_1 for each F_i in $\mathfrak{F}$.
   (b) Based on D_1, D_2, and Map_1, the keywords of each plaintext F_i in $\mathfrak{F}$ can be mapped to a vector with size m_2. Thus, we can establish a stemmed index vector I_2 with dimension m_2 for each F_i in $\mathfrak{F}$. Note that each item in I_2 is a weight of keyword which is the sum of the number of its mapping words in D_1 based on Map_1.
   (c) Based on D_2, D_3, and Map_2 the keywords of each plaintext F_i in $\mathfrak{F}$ can be mapped to a vector with size m_3. Thus, we can establish a synonym index vector I_3 with dimension m_3 for each F_i in $\mathfrak{F}$. Note that each item in I_3 is a weight of keyword which is the sum of the number of its mapping words in D_2 based on Map_2. Therefore, we can obtain a hierarchical index. An example based on Fig. 2 is depicted in Fig. 3.
   (d) Finally, dynamic k-means clustering algorithms can be applied to classify three index vectors $\{I_1\}$, $\{I_2\}$, and $\{I_3\}$.
4. *Encryption.* Data owner executes the following steps to encrypt three index vectors I_1, I_2, and I_3 with private keys SK_1, SK_2, and SK_3.
   (a) To extend the dimension of I_1 from m_1 to $(m_1 + 2)$ and setting the (m_1+1)-th item is a random value ϵ_1 as well as the (m_1+1)-th item is 1.
   (b) To split I_1 into $\{I_1', I_1''\}$ as follows. For each $i_{1j} \in I_1$, $i_{1j}' \in I_1'$, $i_{1j}'' \in I_1''$, and $v_{1j} \in V_1$,

$$i_{1j}' = i_{1j}'' = i_{1j}, \text{ if } v_{1j} = 1 \in V_1 \tag{1}$$

$$i_{1j}' = 1/2 \cdot i_{1j} + \delta_1, \text{ otherwise} \tag{2}$$

$$i_{1j}'' = 1/2 \cdot i_{1j} - \delta_1, \text{ otherwise}, \tag{3}$$

   where δ_1 is a random value.

(c) The ciphertext of I_1 is computed by

$$Enc(I_1) = (\Gamma_1^T \cdot I_1', \Gamma_2^T \cdot I_1''). \tag{4}$$

(d) By the similar approach, the ciphertexts of I_2 and I_3 can be computed by

$$Enc(I_2) = (\Lambda_1^T \cdot I_2', \Lambda_2^T \cdot I_2'') \tag{5}$$

and

$$Enc(I_3) = (\Omega_1^T \cdot I_3', \Omega_2^T \cdot I_3''). \tag{6}$$

Finally, data owner uploads $Enc(I_1)$, $Enc(I_2)$, and $Enc(I_3)$ to cloud.

5. *Trapdoor generation.* In our scheme, we provide three types of search functionalities: *exact*, *derivative word*, and *synonym*. User selects some keywords kw_i and searching types to data owner. According to different types of searching, data owner executes the following steps with private keys to return trapdoors to user.
   (a) Exact search. Based on D_1, user's kw_i can be mapped into a vector with dimension m_1. Then,
      i. To extend the dimension of the vector from m_1 to (m_1+2) and setting the (m_1+1)-th item is 1 as well as the (m_1+1)-th item is a random value ϵ_2. The resulted vector is called Q_1.
      ii. For the previous $m+1$ elements in Q_1, they multiply a random value γ and for the (m_1+2)-th element, it multiples another random value γ'.
      iii. To split Q_1 into $\{Q_1', Q_1''\}$ as follows. For each $q_{1j} \in Q_1$, $q_{1j}' \in Q_1'$, $q_{1j}'' \in Q_1''$, and $v_{1j} \in V_1$,

$$q_{1j}' = q_{1j}'' = q_{1j}, \text{ if } v_{1j} = 0 \in V_1 \tag{7}$$

$$q_{1j}' = 1/2 \cdot q_{1j} + \delta_1', \text{ otherwise} \tag{8}$$

$$q_{1j}'' = 1/2 \cdot q_{1j} - \delta_1', \text{ otherwise,} \tag{9}$$

      where δ_1' is a random value.
      iv. To compute trapdoor

$$T_{w1} = Enc(Q_1) = (\Gamma_1^{-1} \cdot Q_1', \Gamma_2^{-1} \cdot Q_1''). \tag{10}$$

   (b) Derivative words search. Based on D_2 and Map_1, we can obtain a trapdoor

$$T_{w2} = Enc(Q_2) = (\Lambda_1^{-1} \cdot Q_2', \Lambda_2^{-1} \cdot Q_2'') \tag{11}$$

   by the previous approach.

(c) Synonym search. Based on D_3 and Map_2, we can obtain a trapdoor

$$T_{w3} = Enc(Q_3) = (\Omega_1^{-1} \cdot Q_3', \Omega_2^{-1} \cdot Q_3'') \tag{12}$$

by the previous approach.

6. *Search.* Upon receiving the trapdoor and type of search from user, cloud selects the corresponding index and each cluster center to compute relevance score $RelevenceScore$ for each file. For example, in exact search $RelevenceScore$ is computed by

$$RelevenceScore = (\Gamma_1^T \cdot I_1', \Gamma_2^T \cdot I_1'') \bullet (\Gamma_1^{-1} \cdot Q_1', \Gamma_2^{-1} \cdot Q_1''). \tag{13}$$

Finally, cloud sorted files and returns a top-k ranked results according to $RelevenceScore$.

3 Performance Analysis

In this section, we evaluate the performance of our proposed HMASA scheme by Python 2.7 language on Windows 10 server with Core-i7 Processor with 2.6 GHz and 16 GB RAM. We use KOS blog entries [1] in Bag of Words Data Set [2] as our data set, including 3430 documents and 6906 keywords extracted from the documents. Finally, there are 6906 keywords in exact keyword dictionary, 4601 keywords in stemmed keyword dictionary, and 3378 keywords in synonym dictionary. We randomly choose 10 query keywords for test.

In the following Figs. 4 and 5, we compare our HMASA schem with MRSE [3] in term of search time. It is easy to see that HMASA has a better performance than MRSE.

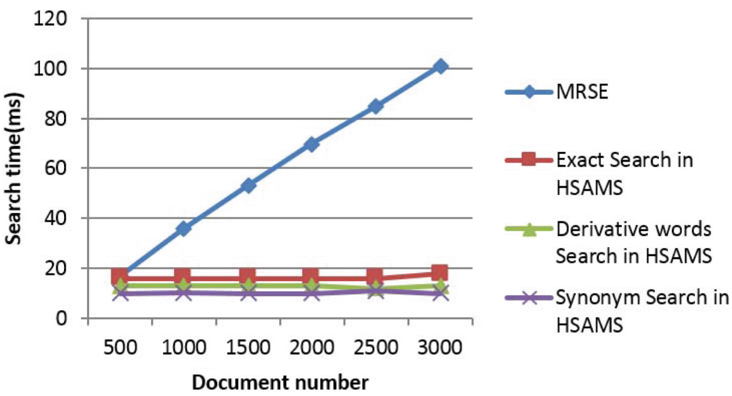


Fig. 4. Search time in different numbers of documents

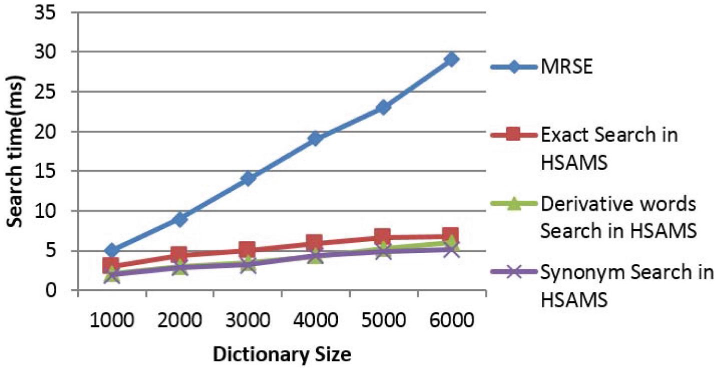


Fig. 5. Search time in different numbers of dictionaries

4 Conclusion

In this paper, a hierarchical semantic approximate multi-keyword ranked search scheme (HSAMS) is proposed. In our scheme, we first take the relationship among different dimensions of index vector into consideration. Performance analysis is demonstrated that our HSAMS scheme has better performance than MRSE.

Acknowledgements. The work of Chien-Ming Chen was supported in part by Shenzhen Technical Project (JCYJ20170307151750788) and in part by Shenzhen Technical Project (KQJSCX20170327161755). The work of Tsu-Yang Wu was supported in part by the Science and Technology Development Center, Ministry of Education, China under Grant no. 2017A13025 and the Natural Science Foundation of Fujian Province under Grant no. 2018J01636.

References

1. https://www.dailykos.com/
2. https://archive.ics.uci.edu/ml/datasets/bag+of+words
3. Cao, N., Wang, C., Li, M., Ren, K., Lou, W.: Privacy-preserving multi-keyword ranked search over encrypted cloud data. IEEE Trans. Parallel Distrib. Syst. **25**(1), 222–233 (2014)
4. Fu, Z., Ren, K., Shu, J., Sun, X., Huang, F.: Enabling personalized search over encrypted outsourced data with efficiency improvement. IEEE Trans. Parallel Distrib. Syst. **27**(9), 2546–2559 (2016)
5. Fu, Z., Wu, X., Guan, C., Sun, X., Ren, K.: Toward efficient multi-keyword fuzzy search over encrypted outsourced data with accuracy improvement. IEEE Trans. Inf. Forensics Secur. **11**(12), 2706–2716 (2016)
6. Fu, Z., Wu, X., Wang, Q., Ren, K.: Enabling central keyword-based semantic extension search over encrypted outsourced data. IEEE Trans. Inf. Forensics Secur. **12**(12), 2986–2997 (2017)

7. Li, H., Liu, D., Dai, Y., Luan, T.H., Shen, X.S.: Enabling efficient multi-keyword ranked search over encrypted mobile cloud data through blind storage. IEEE Trans. Emerg. Top. Comput. **3**(1), 127–138 (2015)
8. Ma, R., Li, J., Guan, H., Xia, M., Liu, X.: EnDAS: efficient encrypted data search as a mobile cloud service. IEEE Trans. Emerg. Top. Comput. **3**(3), 372–383 (2015)
9. Porter, M.F.: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)
10. Song, D.X., Wagner, D., Perrig, A.: Practical techniques for searches on encrypted data. In: Proceedings of the 2000 IEEE Symposium on Security and Privacy, S&P 2000, pp. 44–55. IEEE (2000)
11. Sun, W., Wang, B., Cao, N., Li, M., Lou, W., Hou, Y.T., Li, H.: Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking. IEEE Trans. Parallel Distrib. Syst. **25**(11), 3025–3035 (2014)
12. Wang, C., Cao, N., Li, J., Ren, K., Lou, W.: Secure ranked keyword search over encrypted cloud data. In: 2010 IEEE 30th International Conference on Distributed Computing Systems (ICDCS), pp. 253–262. IEEE (2010)
13. Wang, C., Cao, N., Ren, K., Lou, W.: Enabling secure and efficient ranked keyword search over outsourced cloud data. IEEE Trans. Parallel Distrib. Syst. **23**(8), 1467–1479 (2012)
14. Wu, T.Y., Chen, C.M., Wang, K.H., Pan, J.S., Zheng, W., Chu, S.C., Roddick, J.F.: Security analysis of Rhee et al.'s public encryption with keyword search schemes: a review. J. Netw. Intell. **3**(1), 16–25 (2018)
15. Wu, T.Y., Meng, C., Chen, C.M., Wang, K.H., Pan, J.S.: On the security of a certificateless public key encryption with keyword search. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 191–197. Springer (2017)
16. Wu, T.Y., Meng, C., Wang, K.H., Chen, C.M., Pan, J.S.: Comments on Islam et al.'s certificateless designated server based public key encryption with keyword search scheme. In: International Conference on Genetic and Evolutionary Computing, pp. 199–205. Springer (2017)
17. Wu, T.Y., Meng, C., Wang, K.H., Chen, C.M., Pan, J.S.: Comments on recent proposed Cui et al.'s KASE and Lu et al.'s diBEKS schemes. J. Inf. Hiding Multimed. Sig. Process. **9**(1), 162–196 (2018)
18. Wu, T.Y., Meng, F., Chen, C.M., Liu, S., Pan, J.S.: On the security of a certificateless searchable public key encryption scheme. In: International Conference on Genetic and Evolutionary Computing, pp. 113–119. Springer (2016)
19. Wu, T.Y., Tsai, T.T., Tseng, Y.M.: Efficient searchable ID-based encryption with a designated server. Ann. Telecommun.-annales des télécommunications **69**(7–8), 391–402 (2014)
20. Xia, Z., Wang, X., Sun, X., Wang, Q.: A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data. IEEE Trans. Parallel Distrib. Syst. **27**(2), 340–352 (2016)
21. Zhang, W., Lin, Y., Xiao, S., Wu, J., Zhou, S.: Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing. IEEE Trans. Comput. **65**(5), 1566–1577 (2016)

Mining High Utility Itemsets from Multiple Databases

Jerry Chun-wei Lin[1,2(✉)], Yuanfa Li[1], Philippe Fournier-Viger[3], and Linlin Tang[1]

[1] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen, China
[2] Department of Computing, Mathematics, and Physics, Western Norway University of Applied Sciences, Bergen, Norway
[3] School of Humanities and Social Sciences, Harbin Institute of Technology (Shenzhen), Shenzhen, China
jerrylin@ieee.org, liyuanfa@stu.hit.edu.cn, philfv8@yahoo.com, hittang@126.com

Abstract. In the past, many algorithms have been developed to efficiently mine the high-utility itemsets from a single data source, which is not a realistic scenario since the data may be distributed into varied branches, and the discovered information should be integrated together for making the effective decision. In this paper, we focus on developing an efficient algorithm for synthesizing the mined high-utility itemsets from different sources. A baseline algorithm is first designed and two criteria are then developed to verify whether the designed algorithm is efficient to generate the same number of the high-utility itemsets as the batch-processed algorithm. Experiments are then shown that the designed algorithm has good performance for rule synthesization.

Keywords: Synthesization · High-utility itemsets
Multiple databases · Voted patterns

1 Introduction

Rule synthesization is a major topic in recent decades since the data sources can be distributed into different places in the world; for example, the Walmart retail company holds 11,718 stores in 28 countries. Each branch could produce high volume of transactions every single day. Effective data mining techniques to discover useful knowledge over multiple databases [3,17,19] are essential for the companies having various branches. Thus, it is an interesting topic to merge the multiple data sources from different places for globally decision making. Previous algorithms of data mining mostly focus on mining the interesting information from a single dataset, which is not appropriate for handling the multiple databases. Although many parallel algorithms [2,14–16] were presented to mine the required information, but most of them have the several constraints, for

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 139–146, 2019.
https://doi.org/10.1007/978-3-030-04585-2_17

example, dataset format, inconsistent attributes, or privacy issues. Also, the high cost of the equipments for parallel [2,14–16] and distributed mining [6] is necessary. It is better to mine the interesting patterns individually then synthesize the discovered information for later decision making.

Wu and Zhang [17] designed a weighting model for synthesizing the discovered patterns based on voted model of the frequent rules from different data sources. This model aims at finding the frequent and highly voted rules for centralized decision-making. However, frequent itemset mining only concerns the appearance of the items and discards the other factors, such as important, interesting, unit profit of the items, among others.

High-utility mining [7,11,18] is an emerging topic and many algorithms were respectively presented to speed up the mining performance but non of them focus on developing a synthesizing approach of the discovered high-utility patterns from varied databases. In this paper, we first study the synthesization problem of high-utility itemset mining (HUIM). Major contributions are listed as below.

- A novel algorithm called is first presented to synthesize the discovered HUIs from varied databases.
- Two evolution standards are designed here to verify the performance of the designed approach compared to the conventional algorithm, especially on the target of HUIM.
- Experiments showed that the designed algorithm has nearly the same results as the conventional approach.

2 Review of Related Works

Frequent itemset mining (FIM) and association rule mining (ARM) [1,10] are the basic concepts to mine the relationships between items. Apriori algorithm [1] was first proposed by Agrawal et al., which is used to discover the set of association rules in the level-wise manner based on the minimum support and confidence thresholds. FP-growth [10] was presented to speed up the mining performance by a compressed tree structure. It keeps only the 1-itemsets in the FP-tree structure, and without generation-and-test approach, the k-itemsets can be easily discovered. Several algorithms were respectively presented and applied in varied applications.

The above algorithms mostly focus on mining the required information from a single database. Zhang et al. [20] focused on the analytics of local patterns and defined a pattern called voted patterns in the environment of multiple databases. Wu and Zhang [17] proposed a weighting model for synthesizing global patterns based on voted frequent rules from different the data sources. This model aims at synthesizing highly frequent association rules that is supported or voted for by a large number of data sources. Adhikari [3] then presented two types of frequent itemsets in the multiple databases and proposed an algorithm for synthesizing the discovered information. Several parallel approaches of frequent itemset mining [14–16] were also developed to handle the big data problem in distributed environment but require the highly cost hardware or software for rule maintenance.

High-utility itemset mining (HUIM) [7] is an extension of FIM/ARM but considers more factors such as unit profit and the quantity of the items to discover the set of high-utility itemsets (HUIs). Yao et al. [18] designed several mathematical properties of utility constraints and developed two pruning strategies to reduce the size of the candidates for minin high-utility itemsets. Liu et al. [11] presented a two-phase algorithm to mine the high-utility itemsets, which can show better performance than the previous works, especially in terms of speed and memory cost. Several extensions of HUIs were respectively presented, which may focus on the efficiency problem [5], incremental problem [8,12,21], and the constraint-based problems [9,13]. To the best of our knowledge, non of the works focus on synthesizing the discovered HUIs from multiple databases.

3 Preliminaries and Problem Statement

In this paper, we consider a company with n branches, and each branch has a database D such as: $\{D_1, D_2, \ldots, D_n\}$. Each branch transfer the discovered information to the data center, which is used to synthesize the discovered information for later decision-making. In this paper, we focus on the issue of high-utility itemset mining (HUIM), thus each branch forwards the set of the discovered high-utility itemsets (HUIs) to the data center. Let $I = \{i_1, i_2, \ldots, i_m\}$ be a finite set of m distinct items. For each database D_k, it consists of several transactions such as $D_k = \{T_1, T_2, \ldots, T_p\}$ and each $T_j \in D_k$. Each i_r in T_j is with its purchase quantity such as $q(i_r, T_j)$, and has its own unit profit value as $p(i_r)$. An itemset X consists of several items such that $X = \{i_1, i_2, \ldots, i_q\}$, and it is called as the K-itemset where K is the size of an itemset X such that $K = |X|$. The definitions and problem statement are given as follows.

Definition 1. *The $u(i_r, T_j)$ is used to denote as the utility of an item i_r in a transaction T_j, and defined as:*

$$u(i_r, T_j) = q(i_r, T_j) \times p(i_r) \tag{1}$$

Definition 2. *The $u(X, T_j)$ is used to denote as the utility of X in a transaction T_j, and defined as:*

$$u(X, T_j) = \sum_{i_r \in X \wedge X \in T_j} u(i_r, Tj) \tag{2}$$

Definition 3. *The $u(X)$ is used to denote as the utility of X in the database D, and defined as:*

$$u(X) = \sum_{X \subseteq T_j \wedge T_j \in D} u(X, T_j) \tag{3}$$

Definition 4. *The $tu(T_j)$ is used to denote as the transaction utility of a transaction T_j in a database D, and defined as:*

$$tu(T_j) = \sum_{i_r \in T_j} u(i_r, T_j) \tag{4}$$

Definition 5. *The TU is used to denote as the total utility of a database D, and defined as:*

$$TU = \sum_{T_j \in D} tu(T_j) \tag{5}$$

Definition 6. *An itemset X is called as a high-utility itemset (HUI) if its utility is no less than the minimum utility count of the database, which is defined as:*

$$HUI \leftarrow \{X | u(X) \geq TU \times \delta\}, \tag{6}$$

where δ is the minimum high utility threshold which can be adjusted by users' preference.

For synthesizing the discovered rules from multiple databases, suppose that the discovered HUIs from different databases are $\{H_1, H_2, \ldots, H_n\}$, and each H_k consists of several rules such as $H_k = \{R_1, R_2, \ldots, R_m\}$, where m is the number of HUIs in H_k. Suppose $H' = \{H_1, H_2, \ldots, H_n\}$, with q HUIs. We first find the voted patterns by considering the number of the rules in the data sources. A minimum voted threshold is defined as δ, and the definition is given below.

Definition 7. *A pattern is considered as a highly voted pattern if it holds the following equation as:*

$$Num(R_t) \geq \delta \times n, \tag{7}$$

where n is the number of data sources.

After that, we utilize the weighted model [17] to assign the weight of each rule R_t.

$$w_{R_t} = \frac{Num(R_t)}{\sum_{j=1}^{n} Num(R_j)}, \tag{8}$$

where $t = \{1, 2, \ldots, q\}$, and $Num(R_j)$ is the number of database containing rule R_t.

Based on the above equation, we can obtain the weight of each rule. The weight of the database can then be obtained by the following equation.

$$w_{D_k} = \frac{\sum_{R_t \in H_k} Num(R_t) \times w_{R_t}}{\sum_{k=1}^{n} \sum_{R_h \in S_k} Num(R_h) \times w_{R_h}} \tag{9}$$

We can normalize the values within the interval of [0, 1]. After obtaining the weight of each database, we can obtain the updated total utility in different databases, which can be calculated as:

$$STU = \sum_{k=1}^{n} w_{D_k} \times TU_k. \tag{10}$$

Thus, if the updated utility of an itemset from varied databases is no less than the $STU \times \mu$, it is then considered as a HUI in the multiple databases. Note that the μ is the users' predefined minimum utility threshold which can be adjusted by users' preference. The problem definition is given below.

Problem Statement: The purpose of this paper is to synthesize the discovered HUI from different databases considering weighted model for rule integration. The goal is to find the a pattern is a highly voted pattern and also a high utility pattern compared to the traditional HUIs from the multiple databases.

4 The Designed Algorithm

For the designed algorithm, the weight of the discovered HUI is then calculated, which will be used later to obtain the weight of each database. After that, the weight of each rule is then calculated again based on the weight of different databases, and if a rule is a highly voted-utility pattern, it will be returned as the desired pattern and will be forwarded to the data center for synthesization. The detailed algorithm is then described as follows.

INPUT: $\{D_1, D_2, \ldots, D_n\}$, n databases; $\{TU_1, TU_2, \ldots, TU_n\}$, total utility in n databases; $\{H_1, H_2, \ldots, H_n\}$, the discovered HUIs in different databases forming different rules $\{R_1, R_2, \ldots, R_q\}$; μ, the minimum high-utility threshold; δ, the minimum voted threshold.
OUTPUT: the set of highly voted-utility patterns.
STEP 1: find the voted pattern by eq. (7)
STEP 1: find the weight of each rule by eq. (8)
STEP 2: find the weight of each database by eq. (9)
STEP 4: for the satisfied patterns, find the highly voted-utility pattern if $\sum_{t=1}^{q} u(R_t) \times w_{R_t} \geq STU \times \mu$.
STEP 5: return the results of highly voted-utility patterns.

---

5 Experiments

In this section, we conducted several experiments to evaluate the performance of the proposed algorithm. The HUI-Miner algorithm is then used to compare with the designed algorithm in terms of the size of the discovered high-utility itemsets. Experiments are conducted on a computer with an Intel Core i7-6700 4 GHz processor and 8 GB of the memory, running the 64-bits Microsoft Windows 10 operating system. The characteristics of the four conducted databases [4] are shown in Table 1. Two new evaluation criteria are then stated as follows.

Table 1. The conducted databases in the experiments.

	# of trans.	# of items	Avglen
Chainstore	1,112,949	46,086	7.3
Kosarak	990,002	41,270	8.1
Retail	88,162	16,470	10.3
Accidents	340,183	468	33.8

- **Error count (EC)**: the number of the error results of the HUI, which indicates that an itemset was not a HUI but be the HUI when the rules are merged from multiple databases.
- **Number difference (ND):** the number of different results of the HUI, which indicates that an itemset was a HUI but becomes a non-HUI when the rules are merged from multiple databases.

In the experiments, each data source has the same size of the transactions, and each database is then randomly divided into multiple parts, and each part has the same database size with the others. The minimum utility threshold μ of four databases such as chianstore, kosarak, retail, and accidents are respectively set as: 0.002, 0.02, 0.005, and 0.14. Results are then shown in Table 2.

Table 2. The conducted results.

	Chainstore		Kosarak		Retail		Accidents	
	EC	ND	EC	ND	EC	ND	EC	ND
2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
5	0.00	0.04	0.00	0.00	0.00	0.15	0.00	0.08
10	0.00	0.04	0.00	0.00	0.00	0.12	0.00	0.12

From the above experiment, we can obtain that the designed algorithm could not obtain the EC (always remain stable at 0.00) when the number of the databases are respectively set as 2, 5, and 10. In some cases, when the database is divided into more parts, the ND increases along with the number of multiple data sources, which is reasonable and acceptable. Generally, the proposed algorithm can achieve good performance in terms of EC and ND.

6 Conclusion

In this paper, we first propose a synthesizing model to integrate the discovered high-utility itemsets from multiple databases. Two novel measures are also introduced here to verify the model of the synthesization. Experiments are then conducted on four databases and the results achieve good performance compared to the traditional algorithm used in high-utility itemset mining.

Acknowledgements. This research was partially supported by the Shenzhen Technical Project under JCYJ20170307151733005 and KQJSCX20170726103424709 and by the National Natural Science Foundation of China (NSFC) under grant No. 61503092.

References

1. Agrawal, R. Srikant, R.: Fast algorithms for mining association rules in large databases. In: The International Conference on Very Large Data Bases, pp. 487–499 (1994)
2. Agrawal, R., Shafer, J.C.: Parallel mining of association rules. IEEE Trans. Knowl. Data Eng. **8**(6), 962–969 (1996)
3. Adhikari, A.: Synthesizing global exceptional patterns in different data sources. J. Intell. Syst. **21**(3), 293–323 (2012)
4. Fournier-Viger, P., Lin, J.C.W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H.T.: The SPMF open-source data mining library version 2. In: The Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 36–40 (2016)
5. Fournier-Viger, P., Lin, J.C.W., Duong, Q.H., Dam, T.L.: FHM +: faster high-utility itemset mining using length upper-bound reduction. The International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 115–127 (2016)
6. Gan, W., Lin, J.C.W., Chao, H.C., Zhan, J.: Data mining in distributed environment: a survey. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **7**(6), 1–19 (2017)
7. Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C., Tseng, V.S., Yu, P.: A survey of utility-oriented pattern mining, https://arxiv.org/pdf/1805.10511.pdf (2018)
8. Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C., Hong, T.P., Fujita, H.: A survey of incremental high-utility itemset mining. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. **8**(2), 1–23 (2018)
9. Gan, W., Lin, J.C.W., Fournier-Viger, P., Chao, H.C., Fujita, H.: Extracting non-redundant correlated purchase behaviors by utility measure. Knowl.-Based Syst. **143**, 30–41 (2018)
10. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Discov. **8**(1), 53–87 (2004)
11. Liu, Y., Liao, W.K., Choundhary A.: A fast high utility itemsets mining algorithm. In: The International Workshop on Utility-based Data Mining, pp. 90–99 (2005)
12. Lin, J.C.W., Gan, W., Hong, T.P., Zhang, B.: An incremental high-utility mining algorithm with transaction insertion. Sci. World J. **161564**, 15 pages (2015)
13. Lin, J.C.W., Ren, S., Fournier-Viger, P.: MEMU: more efficient algorithm to mine high average-utility patterns with multiple minimum average-utility thresholds. IEEE Access **14**(8), 7593–7609 (2018)
14. Otey, M.E., Parthasarathy, S., Wang, C., Veloso, A., Meira, W.: Parallel and disctributed methods for incremental frequent itemset mining. IEEE Trans. Syst. Man Cybern. Part B: Cybern. **36**(6), 2439–2450 (2004)
15. Ozkural, E., Ucar, B., Aykanat, C.: Parallel frequent item set mining with selective item replication. IEEE Trans. Parallel Distrib. Syst. **22**(10), 1632–1640 (2011)
16. Sohrabi, M.K., Barforoush, A.A.: Parallel frequent itemset mining using systolic arrays. Knowl.-Based Syst. **37**, 462–471 (2013)

17. Wu, X., Zhang, S.: Synthesizing high-frequency rules from different data sources. IEEE Trans. Knowl. Data Eng. **15**(2), 353–367 (2003)
18. Yao, H., Hamilton, H.J.: Mining itemset utilities from transaction databases. Data Knowl. Eng. **59**(3), 603–626 (2006)
19. Zhong, N., Yao, Y.Y.Y., Ohshima, M.: Peculiarity oriented multidatabase mining. IEEE Trans. Knowl. Data Eng. **15**(4), 952–960 (2003)
20. Zhang, S., Zhang, C., Wu, X.: Knowledge Discovery in Multiple Databases. Springer, London (2004)
21. Zhang, B., Lin, J.C.W., Shao, Y., Fournier-Viger, P., Djenouri, Y.: Maintenance of discovered high average-utility itemsets in dynamic databases. Appl. Sci. **8**(5), 1–19 (2018)

Design Analysis and Control Research of Vehicle Switched Reluctance Motor

Zhong-Shu Liu[1,2](✉)

[1] School of Information Science and Engineering, Fujian University of Technology, Fuzhou 350118, China
lzs@fjut.edu.cn

[2] The Key Laboratory for Automotive Electronics and Electric Drive of Fujian Province, Fujian University of Technology, Fuzhou 350118, China

Abstract. Compared with the widely used vehicle motors currently, such as asynchronous motor, permanent magnet brushless motor and transverse magnetic motor, switched reluctance motor has unparalleled advantages in cost, efficiency, speed regulation performance and reliability, and therefore it has great potential in vehicle designing and manufacturing By using Ansoft Maxwell software a 10 kW three-phase with 12/8 poles switched reluctance motor was optimum designed, its performance characteristics are analyzed and control strategy is summarized in this paper.

Keywords: Vehicle · Switch reluctance motor · Design · Characteristic Simulation

1 Introduction

In recent years, with the global shortage of oil resources and the increasingly serious environmental pollution caused by fuel vehicles, the development of electric vehicles has become the focus in development of global automobile industry. In many developed countries in the world, the research on motor vehicles has been started very early [1]. On the whole, the technology on motors for vehicle in Japan has been very mature, for it has begun to develop from 1960s. Although the United States started lately, it grew rapidly. Electric motors which are used for electric vehicles operate by taking energy from batteries, so they need to be efficient in all kinds of environments. At present, the commonly used motors for vehicle are as the following: DC motor, asynchronous motor, switch reluctance motor and permanent magnet synchronous motor. Among all kinds of vehicle motors, switched reluctance motor (SRM) stands out because of its advantages of good starting performance, wide speed adjustment range, high efficiency, low maintenance cost and high system reliability, so it becomes an ideal vehicle motor recently. By using Ansoft Maxwell software, a three-phase switch reluctance motor with 12/8 poles structure was optimum designed in this paper. Meanwhile, the motor structure and performance parameters was determined. Applying the two-dimensional finite element analysis method to the motor [2], while the relative location of stator pole center line relative to the center line of the rotor pole was changing, the electromagnetic field distribution inside the motor corresponding to the

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 147–155, 2019.
https://doi.org/10.1007/978-3-030-04585-2_18

position was analyzed. Finally, when motor being at low speed the mode of current chopper control was adopted, while it is in high speed the mode of angular position control was adopted. By using these two control modes we can realize the stable operation of the motor under different working conditions, which can meet the requirements of the electric vehicle's frequent starting, acceleration and climbing [3].

2 Design and Analysis of Vehicle Switched Reluctance Motor (SRM)

2.1 The SRM Structure

The switch resistive motor (SRM) used in vehicle is a double salient pole iron core structure. Its stator and rotor are made of ordinary silicon steel laminated. The stator has three-phase, 12 salient poles, each of which is only wound centrally group, so it is easy to manufacture with simple insulation structure. Two radial relative windings of the stator are connected in series to one phase winding. The rotor has 8 poles, with neither winding nor permanent magnet on it, so it has very good mechanical strength, can be used for ultra-high speed operation.

Figure 1 shows the structure of three-phase 12/8 poles SRM. The stator winding is powered by a DC supply, and then connect electronic switches and diodes to form a loop. The reluctance of a switched reluctance motor changes with the electric angle of the center line of stator pole relative to the rotor pole. The operating principle of SRD follows the “magneto-resistive minimum principle”. That is, the flux is always closed along the path with the least resistance [4]. The rotation direction of SRM depends on the electrification sequence of stator winding.

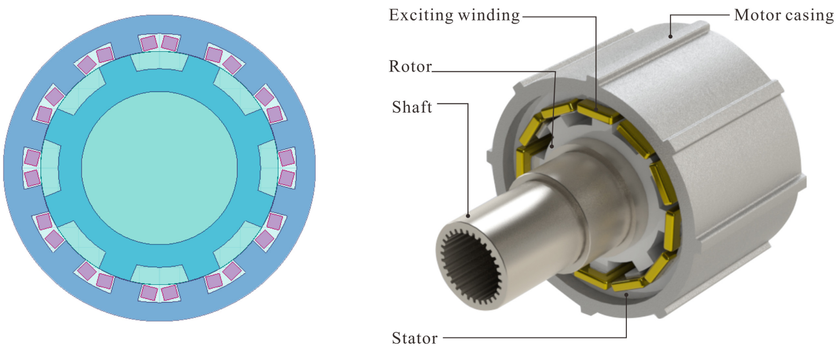


Fig. 1. Structure of three-phase SRM.

2.2 Optimum Design of SRM

The basic idea of the optimum design of the SRM should be to establish an accurate mathematical model firstly, then determine the electromagnetic load and main dimensions, and finally carry out the optimization design [5]. When designing mathematical model by using the nonlinear model of switch reluctance motor, we make the flux fit with vehicle nonlinear function, and then calculate the inductance function corresponding to different position of the stator and rotor pole. The analysis results show that the electromagnetic torque of the switch reluctance motor is directly proportional to the square of the armature current, the direction of the torque has nothing to do with the direction of current, while it is only related to the rate of change of inductance.

When designing vehicle switched reluctance motor we should select lower magnetic load and higher electric load. Magnetic load B can be estimated by the following formula:

$$B \approx (0.3 \sim 0.55)\frac{N_s}{2N_r}B_s \tag{1}$$

In the above formula, B_s is the stator magnetic density, usually we can select 1.3–2.1T, N_s is the number of stator salient pole; N is the number of rotor salient pole.

The selection of electric load can be made by the following formula:

$$A = \frac{2qNI_{eff}}{\pi D_1} \tag{2}$$

In the above formula, N is the turns per phase of stator winding; I_{eff} is the effective value of each phase current of the stator winding, D_1 is the stator internal diameter [6].

The main dimensions of the switched reluctance motor are the specified inner diameter and the effective length of the stator core. When the main dimensions are determined, the weight, the cost and the shape of the motor are also determined [7]. The design steps of main dimension of the motor are shown in Fig. 2.

Finally, Ansoft Maxwell software was used to optimize the design of the motor. The design results are shown in Table 1.

2.3 Finite Element Magnetic Field Analysis of Motor

In this paper, Ansoft Maxwell software is used to make finite element magnetic field analysis to the designed SRM, then the theoretical basis for the nonlinear simulation and control of SRM is provided. Due to the existence of DC source in the solution region, the following assumptions are made during the analysis process [8]. First, the end magnetic field effect of the motor should be ignored, and the magnetic field is uniformly distributed along the axial direction. Second, the magnetic field is confined to the motor interior. Finally, the eddy current effect of alternating magnetic field in conductive material should be ignored. Therefore, the magnetic field of SRM can be treated as a nonlinear steady-state field [9].

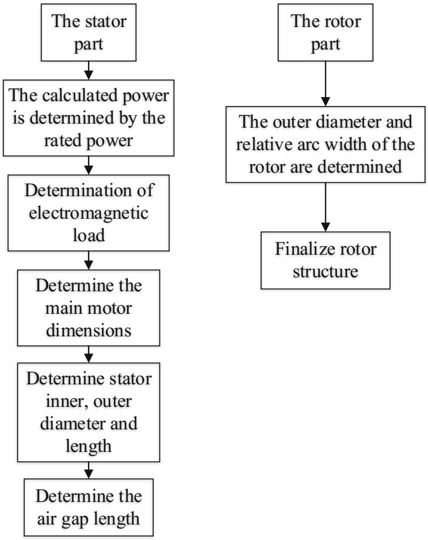


Fig. 2. Block diagram of main dimensions determination of SRM.

Table 1. Structure parameters of the SRM

Parameter	Value	Parameter	Value
The rated voltage	324/V	The rated speed	1500/rpm
The rated power	10 kW	The rotor pole	8
The stator pole	12	Outer diameter of the rotor	235.6/mm
Outer diameter of the stator	310/mm	Inner diameter of the rotor	155/mm
Inner diameter of the stator	236.4/mm	The air gap length	0.4/mm
The stator yoke thickness	19.05/mm	The rotor yoke thickness	22.8/mm
The stator arc	0.45/rad	The rotor arc	0.36/rad
The length of the core	190/mm	The wire diameter	0.95/mm

3 Finite Element Analysis of Magnetic Field

Now we will analyze the SRM by using two-dimensional finite element method, which is a numerical calculation method based on discretization [10]. By this way, firstly we use Maxwell 2D software package for grid subdivision mesh subdivision which is the most crucial step to finite element discretization. Figure 3 is the grid subdivision diagram of SRM. When the center line of the stator pole and center line of the rotor slot coincide, the angle is 0°. While the center line of the stator pole and the center line of the rotor polar coincide, the angle is 45°.

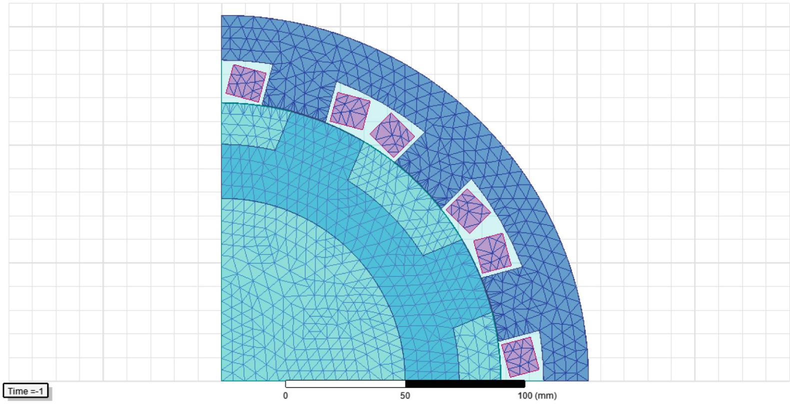


Fig. 3. The grid subdivision diagram of SRM.

Figure 4 shows the flux density cloud diagram of the switched reluctance motor when rotor angle is 15°. It can be seen from the figure that the magnetic flux leakage phenomenon is relatively serious at this time, and the magnetic circuit is in an unsaturated state, maximum magnetic density is 1.621T.

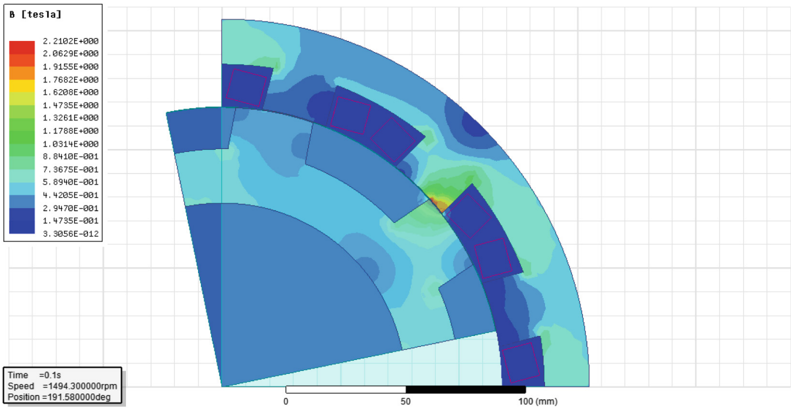


Fig. 4. Flux density cloud diagram when rotor angle is 15°.

When the rotor rotate at 0°–150° electric angle, the problem of magnetic field analysis can be analyzed separately. Here, Flux linkage picture of the distribution of magnetic field lines correspond to different electric angles were shown in Fig. 5. The magnetic flux of SRM consists of main magnetic flux and leakage magnetic flux [11]. When the angle is 0°, the gap magnetic resistance is biggest, and the magnetic flux leakage is largest. When the rotor position angle gradually increases, the main magnetic flux and leakage magnetic flux both increase. When the center line of the stator and the center line of the rotor pole coincide, the flux of the motor is mainly the main magnetic flux. And the saturation degree of stator yoke and stator pole is very serious [12]. At this point, the maximum magnetic density value in stator-yoke flux density is 1.011T, stator-pole flux density is 0.737T, rotor-pole flux density is 0.846T, rotor-yoke flux density is 0.616T, which can fully meet the design requirements.

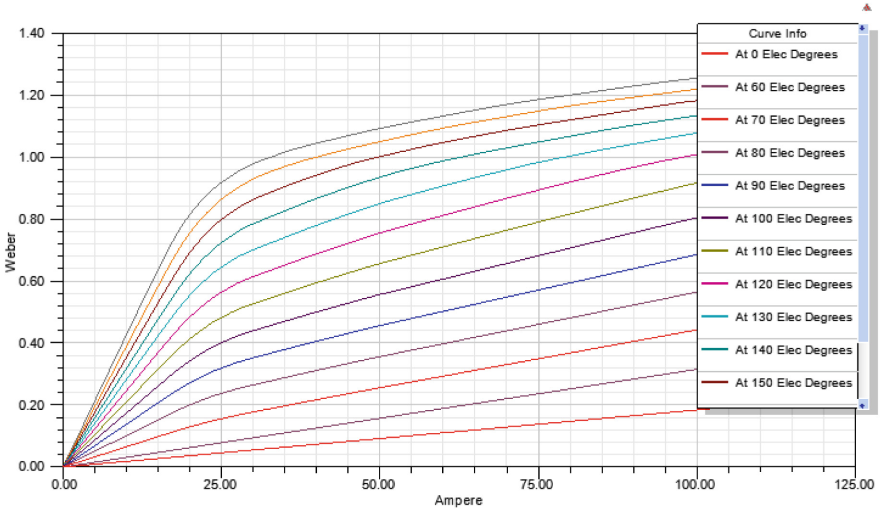


Fig. 5. The flux linkage diagram at different electric angles of SRM.

4 Control Strategy and Simulation Results

4.1 Control Strategy

Vehicle running characteristics of switched reluctance motor can be divided into three regions: the constant torque region (low speed region), the constant power region (high speed region) and the nature region (series excitation region). Under normal circumstances, the SRM runs in the first two regions. In the constant torque region, the motor speed is low, while the counter electric motive force is also very small, it needs to control the chopping current, so the mode of current chopping control (CCC mode) is adopted. In the constant power region, the motor has a higher speed and it leads to a higher counter electromotive force. At this point, the motor speed can be controlled by adjusting the opening angle and off angle of the main switch tube that is by changing the switch phase corresponding to the rotor position, which is called the angle position control mode (APC mode) [13].

Traditional control methods, such as PID-controlled switched reluctance motor drive system, are widely used on some occasions, but it cannot effectively overcome changes caused by transmission objects and load parameters over the wide range as well as the influence of nonlinear factors on the system [14]. Therefore, it is difficult to meet the requirements of different operating conditions such as frequent starting, acceleration and deceleration of electric vehicles. A kind of mode applicable to vehicle motor compound control is introduced in this paper, which means we have the mode of fuzzy slip mode control (FSMC) together with CCC mode and APC mode, it is a new complicated control mode. FSMC is a kind of relatively advanced control mode, which combines the advantages of fuzzy logic control (FLC) and slid mode control (SMC) [9]. The control mode of FSMC for SRM is composed of double closed loop control system, one is the internal current loop, and the other is outside speed loop. The FSMC control system of the switched reluctance motor is shown in Fig. 6.

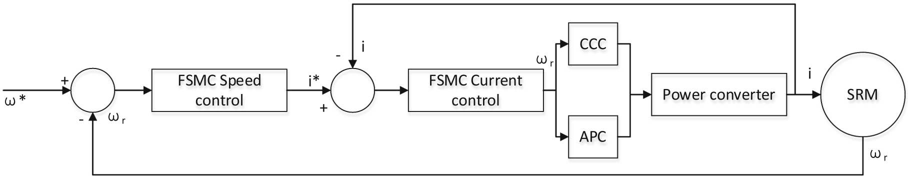
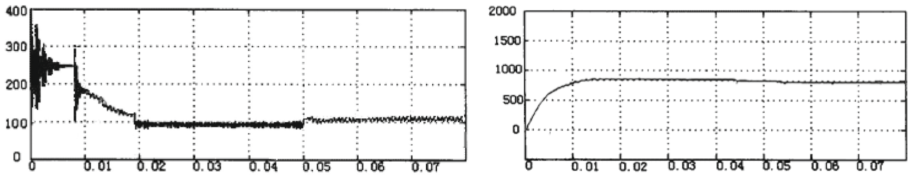


Fig. 6. SRM FSMC control system.

4.2 Simulation Results

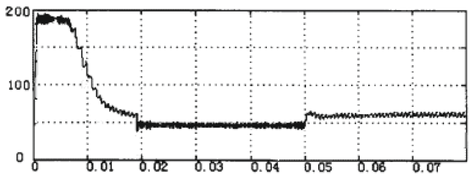
In this paper, the control system of the SRM is modeled and simulated under MATLAB/Simulink. The power supply voltage is 324 V DC voltage. When the given speed of the system is 700 r/min and1500 r/min, the torque and speed characteristic of the SRM simulation results are shown in Figs. 7 and 8 respectively.



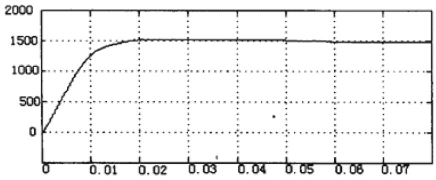
a) Torque characteristics

b) Speed characteristics

Fig. 7. Speed 700 r/min.



c) Torque characteristics



d) Speed characteristics

Fig. 8. Speed 1500 r/min.

It can be seen from Fig. 7 that the system has fast response speed and stable starting process when running at low speed. While running at high speed as shown in Fig. 8, the torque ripple is small and the vehicle runs comfortably and smoothly [15].

5 Conclusion

By using Ansoft Maxwell software, a 10 kw three-phase with 12/8 poles switch reluctance motor (SRM) was optimum designed in this paper. Meanwhile, the dynamic characteristic control strategy and simulation was analyzed. The design and the simulation results show that the designed switch reluctance motor could remain good control characteristic in the wide power and speed range, which is suitable for electric vehicle under various conditions. The research and development of vehicle switching reluctance motor has important practical significance and popularization value in electric vehicle industry.

Acknowledgment. This research was supported by the Science and Technology Program Foundation of Fuzhou Science and Technology Bureau (Grant No. 2016-G-57).

References

1. Zhang, Z., Quan, L., Zhu, X., Zhangu, C., Liu, X.: Design and control of switched reluctance motor used for micro electric vehicles. Micromotors **41**(3), 19–23 (2013)
2. Hung, M.-H., Chen, Q., Chen, Y.: A review of handling continuous and unknown attribute values of C4.5 algorithm. J. Netw. Intell. **3**(1), 1–8 (2018)
3. Liu, Z.: Structure of features and simulation research of the vehicle switched reluctance motor. J. Minjiang Univ. **34**(2), 54–57 (2013)
4. Shih, S.-T., Kam, L.: Multi-fractal texture segmentation for off-road robot vision application. J. Netw. Intell. **3**(3), 189–194 (2018)
5. Wang, M., Chen, Q., Hu, C.: Switched Reluctance Motor Self-Optimizing Control Simulation. Small Spec. Electr. Mach. **43**(12), 77–80 (2015)
6. Chun-Ming, W., Gong, H.-Q., Yang, J.-H., Song, Q.-H., Wang, Y.-J.: An improved FOA to optimize GRNN method for wind turbine fault diagnosis. J. Inf. Hiding Multimed. Signal Process. **9**(1), 1–10 (2018)
7. Chen, S.-H., Chu, Y.-Y.: Enhanced error-recovery CAN bus system using reed-solomon codec. J. Netw. Intell. **2**(4), 220–330 (2017)

8. Pan, J.-S., Nguyen, T.-T., Dao, T.-K., Pan, T.-S., Chu, S.-C.: Clustering formation in wireless sensor networks: a survey. J. Netw. Intell. **2**(4), 1–8 (2017)
9. Lin, S., Huang, X.: Power loss analysis of AC contactor at steady closed state with electromagnetic-thermal coupling method. J. Inf. Hiding Multimed. Signal Process. **8**(2), 290–299 (2017)
10. Yu, M., Du, Z., Liu, X., Qun, D., Chen, H.: The method of obtaining best unary polynomial for the chaotic sequence of image encryption. J. Inf. Hiding Multimed. Signal Process. **8**(5), 1103–1110 (2017)
11. Li, J., Oiu, J., Dong, X.: Development of switched reluctance motor control system for vehicles based on model design. Micromotors **43**(7), 82–84 (2015)
12. Wang, J., Sun, J., Zheng, Z.: Position control of a fault-tolerant direct-drive hub motor used by electric vehicle. Control Decis. **32**(16), 1133–1136 (2013)
13. Tan, C., Wang, H., Chen, L.: Finite-element modal analysis of stator vibration for BSRM with double windings. J. Hohai Univ. **45**(1), 90–94 (2017)
14. Gan, S., Wang, J., Song, X., Tao, T., Tang, J.: Direct instantaneous of SRM for vehicle based on modified TSF. Small Spec. Electr. Mach. **45**(7), 49–53 (2017)
15. Zhou, Z., Wang, H., Yang, Z.: Research on SRM direct instantaneous torque control System with single neuron PID speed controller. Mach. Build. Autom. **45**(2), 173–176 (2016)

Intelligent Computing Techniques and Its Applications (Invited Session 02)

Modified Image Dehazing Method Based on Dark Channel Prior

Junpeng Hu[1,2], Xinrong Cao[2], Xinwei Chen[2], Zuoyong Li[2(✉)], and Fuquan Zhang[2]

[1] College of Mathematics and Computer Science, Fuzhou University, Fuzhou, China
hujunpengfzu@126.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China
cxrxmu@163.com, chen_xinwei@126.com, fzulzytdq@126.com, 8528750@qq.com

Abstract. Image dehazing is an important research topic in the field of image processing and computer vision. Image dehazing aims to remove haze in images and make image scenes clearer. Image dehazing based on dark channel prior is a currently popular type of methods. However, image dehazing results obtained by existing methods based on dark channel prior usually have color distortion and low brightness causing partial image details invisible. To alleviate this issue, we presented a modified method based on dark channel prior. First, the proposed method estimates the value of atmospheric light by using a quadtree algorithm, and uses the dark channel prior to pixelwisely estimate and optimize medium transmission. Second, the proposed method uses a classic atmospheric scattering model to generate an initial image dehazing result, and transforms the result from RGB (Red, Green, and Blue) color space to HSV (Hue, Saturation, and Value) color space. Finally, the proposed method conducts the CLAHE (Contrast Limited Adaptive Histogram Equalization) algorithm on the V component of the initial result, and maps the result into the RGB space to obtain final image dehazing result. Experimental results showed that the proposed method effectively alleviated color distortion and made image scenes clearer in image dehazing results.

Keywords: Image dehazing · Dark channel prior · HSV · Color distortion

1 Introduction

Image dehazing (i.e., image haze removal) is an important research topic in the field of image processing and computer vision. As compared with haze-free images, haze images are degraded by fog or haze, and they usually have low image contrast, low visibility, and low color fidelity. This degradation significantly affects real applications of haze images and videos in military, monitoring, traffic, target tracking, and automatic navigation. Image dehazing aims to remove the haze and make image scene clearer. To remove the image haze, researchers [1–5] proposed some image dehazing methods in recent years. Image dehazing methods can be roughly divided into two categories, i.e., image enhancement based methods and image restoration based methods.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 159–169, 2019.
https://doi.org/10.1007/978-3-030-04585-2_19

Image enhancement based methods didn't consider mechanism of haze images' degradation, and directly enhanced haze images to improve image visual effect. This type of methods can be further divided into global methods [2, 3] and local methods [4, 5]. Global enhancement methods achieved image dehazing according to statistical information of the whole image. And conversely, local enhancement methods achieved image dehazing according to local image statistical information.

Image restoration based methods studied mechanism of haze image degradation, constructed image degradation models to obtain image dehazing results. This type of methods can be further divided into depth relationship based methods [6, 7] and prior information based methods [8–11]. Depth relationship based methods first estimated image depth, then solved parameters of image degradation model, and finally brought parameters into the model to obtain image dehazing result. This type of methods had strong pertinence, and had high requirement on computational hardware, which greatly restricted their real applications. Prior information based methods usually obtained better image dehazing effects among various methods, because they considered variation of fog concentration in real images. For example, He et al. [9] proposed a classic dehazing method based on dark channel prior, and the method is briefly called CDDCP in this paper. The CDDCP method [9] was motivated by abundant statistical analysis on real haze images and haze-free images. Jiang et al. [10] presented an improved dehazing method based on dark channel prior, and the method is briefly called IDDCP in this paper. The IDDCP method [10] designed a tolerance correction mechanism to eliminate color distortion in sky region of image dehazing results obtained by CDDCP. He et al. [11] proposed a guided filter to instead of soft matting used in CDDCP, which dramatically improved image dehazing speed. Yang et al. [12] also introduced the guided filter into adaptive image dehazing.

After exploring existing image dehazing methods based on dark channel prior, we found that they still have two limitations. One limitation is that their image dehazing results usually have low brightness, which causes partial image details invisible. The other limitation is that their image dehazing results usually have color distortion. To alleviate these issues, we proposed a modified dehazing method based on dark channel prior. Experimental results on several real representative haze images demonstrated superiority of the proposed method over the counterparts.

2 Classic Image Dehazing Based on Dark Channel Prior

2.1 Atmospheric Scattering Model

McCartney [13] first proposed a famous atmospheric scattering model to describe the principle of generating a digital haze image, and the model is defined as

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

where I is the observed haze image, J is the ideal haze-free image, t is the medium transmission describing the portion of the light that is not scattered and reaches the camera. Higher t value indicates stronger ability of the light penetrating fog or haze. A

is the atmospheric light (i.e., the illumination intensity at the infinite far region along with the line of the observer), which is often hypothesized as a global constant. The goal of image dehazing is to estimate the haze-free image, J, in Eq. (1). To obtain J from I, we need calculate A and t.

2.2 Dark Channel Prior

After abundant statistical analysis on outdoor haze-free images, He et al. [9] first concluded dark channel prior. That is, in most non-sky patches of haze-free images, at least one color channel has some pixels whose intensities are close to zero. The minimum intensity in such a patch is close to zero. The dark channel prior of I is defined as

$$I^{dark}(x) = \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} I^c(y) \right), \tag{2}$$

where I^c is the c-th color channel of I, $\Omega(x)$ is the local patch centered at x. Similarly, we can define the dark channel of the haze-free image J as J^{dark}. For the haze-free image J, except for the sky region, the intensity of J's dark channel is low and tends to be zero, i.e., $J^{dark} \cong 0$.

2.3 Dehazing Based on Dark Channel Prior

2.3.1 Calculation of the Atmospheric Light

Tan's method [8] took the highest image brightness as A. However, once a haze image has sky region or white subject, which easily causes inaccurate estimation of Tan's method on A. To more accurately estimate A, the CDDCP method presented by He et al. [9] used the dark channel image (i.e., I^{dark}) to alleviate the limitation of Tan's estimation on A. Specifically, after sorting all image pixels' intensities in the dark channel image with descending order, the front 0.1% pixels were first chosen as candidate pixels, and then the intensity value of the brightest candidate pixel was taken as A.

2.3.2 Calculation of the Transmission

After obtaining A in Eq. (1), both sides of Eq. (1) are divided by A, and then two minimum filtering are performed on both sides. Thus, Eq. (1) can be written as

$$\min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} \frac{I^c(y)}{A^c} \right) = \tilde{t}(x) \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} \frac{J^c(y)}{A^c} \right) + 1 - \tilde{t}(x). \tag{3}$$

Since the CDDCP method assumed $t(x)$ in Eq. (1) as a constant $\tilde{t}(x)$ in the local patch centered at x, it can be put on the outside of the min operators. According to the definition of the dark channel in Eq. (2) and the assumption that A is a constant, the first item on the right side of Eq. (3) can be reshaped as $\frac{\tilde{t}(x)}{A^c} J^{dark}(x)$. Since J is a haze-free image, $J^{dark}(x) \cong 0$ according to the principle of the dark channel prior. Thus, the first item on the right side of Eq. (3) should be zero. In this case, the transmission $\tilde{t}(x)$ can be calculated by rewriting Eq. (3) as

$$\tilde{t}(x) = 1 - \omega \min_{y \in \Omega(x)} \left(\min_{c \in \{r,g,b\}} \frac{I^c(y)}{A^c} \right) \tag{4}$$

where the added parameter $\omega(0 < \omega \leq 1)$ is used to adaptively keep more fog for distant objects. Generally, the value of ω is positively related with fog concentration.

2.3.3 Image Dehazing

Once the estimated $\tilde{t}(x)$ and A were brought into Eq. (1), and the image dehazing result can be calculated by

$$J(x) = \frac{I(x) - A}{\max\{\tilde{t}(x), \tilde{t}_0\}} + A, \tag{5}$$

where $\tilde{t}_0$ is an added parameter, which is a preset lower bound used to preserve a small amount of haze in very dense haze regions.

3 The Proposed Method

The classic dehazing method based on dark channel prior, which is briefly called CDDCP [9], easily generates color distortion in image bright regions such as sky. The IDDCP method [10] designs a tolerance correction mechanism to eliminate color distortion in sky region. However, image dehazing results obtained by CDDCP [9] and IDDCP [10] usually have low brightness, which causes partial image details invisible and increases visual effect. To alleviate this issue, we proposed a modified image dehazing method in this paper. The proposed method first speeds up estimation of the atmospheric light by a quadtree scheme, then simplifies and optimizes estimation of the transmission to obtain an initial image dehazing result, and finally adaptively increases image brightness to refine the initial image dehazing result.

3.1 Estimation of the Atmospheric Light

Tan's method [8] regards the highest image brightness as the atmospheric light, A. The CDDCP method [9] first uses dark channel image to find candidate pixels, and then regards the intensity value of the brightest candidate pixel as A. In summary, existing methods usually use bright pixels to estimate A. In the proposed method, we used a quadtree algorithm to speed up estimation of A. Specifically, an input image is first uniformly divided into four rectangle regions, and the region with the highest average brightness is chosen to be further uniformly divided into four smaller regions. This process is repeated until the size of the chosen rectangle region smaller than a given threshold. The above estimation process is visually shown in Fig. 1, where the red rectangle region is the finally chosen image region, and its average intensity is taken as the atmospheric light value. Estimation based on the quadtree algorithm has a high convergence speed, thus saving estimation time.

Fig. 1. Estimation of the atmospheric light by using the quadtree algorithm.

3.2 Calculation of Simplified Transmission

The CDDCP method [9] performed twice minimum filtering when calculating the transmission, which easily causes block effect on image edges with large scene depth changes. In addition, the minimum filtering also decreases intensities of most image pixels, thus causing low brightness. Therefore, a constant $\varepsilon_0 (0 < \varepsilon_0 \leq 1)$ is used to replace the first minimum filtering in Eq. (4), and Eq. (4) is rewritten as

$$\bar{t}(x) = 1 - \omega\varepsilon_0 \min_{c \in \{r,g,b\}} \frac{I^c(y)}{A^c}, \tag{6}$$

Let $\varepsilon = \omega\varepsilon_0$, we can deduce that

$$\bar{t}(x) = 1 - \varepsilon \min_{c \in \{r,g,b\}} \frac{I^c(y)}{A^c}, \tag{7}$$

where $0 < \varepsilon \leq 1$, the higher ε is, the stronger the image dehazing effect. However, if ε is too large, the image dehazing result will have low brightness, losing more image details. And conversely, if ε is too small, the image dehazing effect is not evident. Based on abundant experiments, ε is empirically set to 0.8.

Calculation of the transmission in Eq. (7) is pixelwise, and the estimated transmission is enough precise as illustrated in Fig. 2(d). Hence, the proposed method doesn't use the soft matting algorithm [9, 10] or the guided filtering [11] to refine the transmission estimation. As compared with the image dehazing methods [9–11], our estimation for the transmission saves the time performing the minimum filtering, the soft matting, or the guided filtering. Figure 2 exhibits the transmission estimation results obtained by CDDCP [9], IDDCP [10], and the proposed method. In the literature [11], He et al. presented a guided filter, which can be used to replace the soft matting in the process of image dehazing [9] and achieve similar even better image dehazing results. Furthermore, the guided filter costs much smaller time as compared with the soft matting. Therefore, experimental results of CDDCP were obtained by using the guided filter. From Fig. 2, one can observe that the estimated transmission map obtained by the proposed method preserves more image details than other two methods, such as handrails.

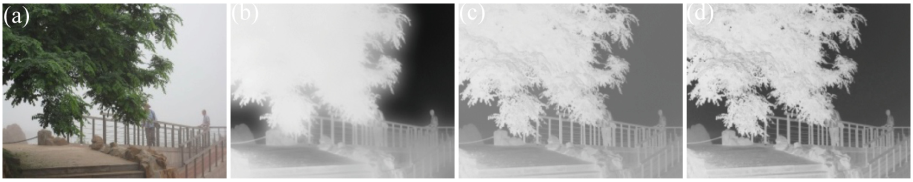


Fig. 2. Estimated transmission maps obtained by different methods: (a) original image, (b) CDDCP [9], (c) IDDCP [10], (d) the proposed method.

3.3 Optimization of Simplified Transmission

Estimation of the simplified transmission solves the problem of block effect. However, it still easily causes color distortion in large bright regions (e.g., sky, white subject, and water surface) of image dehazing results. To resolve this issue, we further optimize the transmission estimation. Specifically, Eq. (3) can be written as

$$\hat{t}(x) = \frac{1 - \min_{y\in\Omega(x)}\left(\min_{c\in\{r,g,b\}} \frac{I^c(y)}{A^c}\right)}{1 - \min_{y\in\Omega(x)}\left(\min_{c\in\{r,g,b\}} \frac{J^c(y)}{A^c}\right)}. \tag{8}$$

After ignoring the hypothesis of dark channel prior and following the operation of replacing the first minimum filtering with a constant $\varepsilon_0 (0 < \varepsilon_0 \le 1)$ in Eqs. (6) and (8) can be rewritten as

$$\hat{t}_0(x) = \frac{1 - \varepsilon_0 \min_{c\in\{r,g,b\}} \frac{I^c(y)}{A^c}}{1 - \varepsilon_0 \min_{c\in\{r,g,b\}} \frac{J^c(y)}{A^c}}. \tag{9}$$

In bright image regions, $\varepsilon_0 \min_{c\in\{r,g,b\}} \frac{J^c(y)}{A^c}$ dose not be close to 0. The denominator in Eq. (9) is smaller than 1, and the actual transmission $\hat{t}_0(x)$ is higher than the transmission $\hat{t}(x)$ estimated according to the dark channel prior, because the denominator should be 1 in Eq. (8) according to the assumption of the dark channel, i.e., $J^{dark} \cong 0$. Therefore, the proposed method introduces a tolerance correction mechanism [10] to improve image dehazing effect on bright image regions such as sky. Specifically, a parameter K is introduced as a tolerance factor. Image regions with $|I(x) - A|$ lower than K is regarded as bright regions, and the transmissions are recalculated. Image regions with $|I(x) - A|$ higher than K is regarded as the regions meeting with the dark channel prior, and the transmission calculation follows the literature [9]. In this case, Eq. (5) is redefined as

$$J(x) = \frac{I(x) - A}{\min\left(\max\left(\frac{K}{|I(x) - A|}, 1\right) \cdot \max\left(\hat{t}(x), \tilde{t}_0\right), 1\right)} + A. \tag{10}$$

Figure 3 shows image dehazing results obtained by CDDCP [9], IDDCP [10], and the proposed method. From Fig. 3, one can observe that the sky region in the dehazing result obtained by CDDCP has evident color distortion and residual fog. The dehazing results obtained by IDDCP [10] and the proposed method have not color distortion and residual fog. However, the IDDCP method [10] refined the transmission estimation by the soft matting algorithm, which costs far longer time than the proposed method.

Fig. 3. Image dehazing results obtained by different methods: (a) original image, (b) CDDCP [9], (c) IDDCP [10], (d) the proposed method.

3.4 Refinement of Image Dehazing Result

Although the proposed method can eliminate color distortion and residual fog in image dehazing results, the dehazing results shown in Fig. 3(d) and Fig. 4(b) still have low brightness, which causes partial image details invisible. To improve image visibility, the proposed method introduces a contrast limited adaptive histogram equalization method briefly called CLAHE [4] to adjust image brightness and enhance image contrast. Specifically, the initial image dehazing result obtained by the proposed method is first transformed from the RGB color space to the HSV color space. Then, the CLAHE algorithm is performed on the V component. Finally, the processed result is transformed from the HSV space to the RGB space as the final dehazing result. The refinement step enhances image brightness and image contrast, thus improving image visibility significantly.

Fig. 4. Initial image dehazing result obtained by the proposed method: (a) original image, (b) the image dehazing result.

4 Experimental Results

To validate superiority of the proposed image dehazing method over the counterparts including CLAHE [4], CDDCP [9], and IDDCP [10], three representative haze images are chosen as testing samples. Three images have the following characteristics: (1) they are from different image scenes. In detail, the first image is captured in a wood scene, the second image is captured in a scene on the top of a mountain, and the third image is captured in a city scene. (2) They have abundant image details, such as the subjects in the wood of the first image, the leaves of the tree in the second image, and buildings in the third image. Therefore, they can be used to observe image dehazing on image details. (3) They are color images, which can be used to observe color distortion in image dehazing results. (4) The first image has not sky region, and the latter two images have large sky regions, which can be used to observe color distortion in sky regions. The convolution kernel used in CDDCP and IDDCP are set to 15×15. Other parameters used in three existing methods are set according to their own literatures. All experiments were performed on a PC with 2.4G CPU and 6G RAM.

Image dehazing results were qualitatively and quantitatively evaluated. Image contrast is used to quantitatively measure image dehazing effect. The image contrast, σ, is the standard deviation of image pixels' gray values, which reflects the overall gray distribution of the image. The image contrast is defined as

$$\sigma = \sqrt{\frac{1}{M \times N} \sum_{i=0}^{M} \sum_{j=0}^{N} (g(i,j) - u)^2}, \tag{11}$$

where $M \times N$ is the image size, $g(i, j)$ is the gray value of the image pixel at i-th row and j-th column, and u is the gray mean of the image. High image contrast value usually indicates good image dehazing effect.

To qualitatively compare image dehazing effects of four methods, Fig. 5 shows image dehazing results of the three haze images obtained by CLAHE [4], CDDCP [9], IDDCP [10], and the proposed method. From Fig. 5, one can observe that image dehazing results obtained by CLAHE have color distortion, because the results are too white, looking like snow scenes. Image dehazing results obtained by CDDCP look dark, which causes partial image details invisible, such as the image details hidden in the wood of the first image, leaves of the tree in the second image, and the building details in the third image. In addition, the image dehazing results on the latter two images have obvious color distortion in sky regions. As compared with CDDCP, IDDCP eliminate the color distortions in sky regions of the latter two images. However, its image dehazing results look more darker than the results obtained by CDDCP, which causes more image details invisible. As compared with the counterparts, the image dehazing results obtained by the proposed method not only eliminate color distortion especially in sky regions, but also look brighter and make more image details visible.

Fig. 5. Qualitative comparison of image dehazing results obtained by different methods: (a) original image, (b) CLAHE [4], (c) CDDCP [9], (d) IDDCP [10], (e) the proposed method.

Quantitative comparisons of image dehazing results are listed in Tables 1 and 2, respectively. Table 1 lists image contrast values of image dehazing results, and Table 2 lists running time of image dehazing results. Image contrast is used to measure image dehazing effect. Running time is used to reflect speed of image dehazing. From two tables, one can observe that: (1) the proposed method obtains the highest image contrast value on the front two images, and obtains the second highest image contrast value on the last image. High image contrast values usually indicate better image dehazing effect. (2) The proposed method costs the least running time on image dehazing, showing the highest image dehazing speed.

Table 1. Image contrast values of image dehazing results

Image number	CLAHE [4]	CDDCP [9]	IDDCP [10]	Proposed
1	44.30	29.17	28.88	**47.72**
2	69.01	59.93	78.58	**79.47**
3	60.94	46.74	**69.34**	63.42

Table 2. Running time (second) of image dehazing results

Image number	CLAHE [4]	CDDCP [9]	IDDCP [10]	Proposed
1	1.00	0.80	75.35	**0.64**
2	1.70	1.48	171.43	**1.29**
3	1.02	1.32	69.35	**0.87**

5 Conclusion

Image dehazing is an important research topic in the field of image processing and computer vision. It aims to remove the haze in a haze image for making the image scene clearer. After exploring the limitations of existing image dehazing methods based on dark channel prior, a modified method is presented in this paper. Contributions of the proposed method are as follows: (1) it simplifies and optimizes estimation of the transmission, improving the estimation accuracy. (2) It applies a contrast limited adaptive histogram equalization algorithm to the V component in the HSV color space to refine image dehazing results. Experimental results on several representative haze images demonstrated that the proposed method alleviates color distortion in image dehazing results, and enhances image brightness and image contrast, which makes more image details visible. In addition, the proposed method has a high speed of image dehazing, making it possibly used in real-time image dehazing applications.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China (61772254), National Nature Science Foundation of China (Grant number: 61871204), Fujian Provincial Leading Project (2017H0030 and 2018H0028), Key Project of College Youth Natural Science Foundation of Fujian Province (JZ160467), Fuzhou Science and Technology Project (2018-S-123 and 2016-S-116), New Century Excellent Talents in Fujian Province University (NCETFJ), Project of Minjiang University (MYK17021), and Major Project of Sichuan Province Key Laboratory of Digital Media Art (17DMAKL01).

References

1. Wu, D., Zhu, Q.: The latest research progress of image dehazing. Acta Automatica Sinica **41**(2), 221–239 (2015)
2. Kim, S.E., Jeon, J.J., Eom, I.K.: Image contrast enhancement using entropy scaling in wavelet domain. Sig. Process. **127**, 1–11 (2016)
3. Eunsung, I., Sangjin, K., Wonscok, K.: Contrast enhancement using dominant brightness level analysis and adaptive intensity transformation for remote sensing image. IEEE Geosci. Remote Sens. Lett. **10**(1), 62–66 (2013)
4. Zhou, W., Liao, H.: Fog image enhancement algorithm based on high frequency and CLAHE. Video Eng. **34**(7), 38–40 (2010)
5. Kaur, A., Singh, C.: Contrast enhancement for cephalometric images using wavelet-based modified adaptive histogram equalization. Appl. Soft Comput. **51**, 180–191 (2017)
6. Narasimhan, S.G., Nayar, S.K.: Vision and the atmosphere. Int. J. Comput. Vis. **48**(3), 233–254 (2002)
7. Kopf, J., Neubert, B., Chen, B., et al.: Deep photo: model-based photograph enhancement and viewing. ACM Trans. Graph. **27**(5), 1–10 (2008)
8. Tan, R.T.: Visibility in bad weather from a single image. In: Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, pp. 1–8. IEEE (2008)
9. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2011)
10. Jiang, J., Hou, T., Qi, M.: Improved algorithm on image haze removal using dark channel prior. J. Circ. Syst. **16**(2), 7–12 (2011)

11. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. **35**(6), 1397–1409 (2013)
12. Yang, Y., Bai, H., Wang, F.: Single image adaptive defogging algorithm based on guidance filtering. Comput. Eng. **42**(1), 265–271 (2016)
13. McCartney, E.J.: Optics of the Atmosphere: Scattering by Molecules and Particles, pp. 23–32. Wiley, New York (1976)
14. Gibson, K.B., Vo, D.T., Nguyen, T.Q.: An investigation of dehazing effects on image and video coding. IEEE Trans. Image Process. **21**(2), 662–673 (2011)

Research on Stage Creative Scene Model Generation Based on Series Key Algorithms

Fuquan Zhang[1,2(✉)], Gangyi Ding[2], Linjuan Ma[3], Yumeng Zhu[2], Zuoyong Li[1], and Lin Xu[4]

[1] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, People's Republic of China
8528750@qq.com, fzulzytdq@126.com

[2] Digital Performance and Simulation Technology Laboratory, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
dgy@bit.edu.com, lyxsmolly@163.com

[3] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
923785608@qq.com

[4] Key Laboratory of Nondestructive Testing, Fuqing Branch of Fujian Normal University, Fuzhou 350300, People's Republic of China
71471418@qq.com

Abstract. This paper proposes a method to construct a stage scene systematically. We propose how to build key algorithms for each element of the stage based on this method. The scene generation model is constructed to generate the creative stage scene we need based on the combination of these algorithms. We will start with 3D model adversarial generation, style migration generation, scene construction layout, scheme evaluation and scheme selection and so on. Then we will elaborate GANs algorithm, style transfer network, spatial combination and other algorithms.

Keywords: Creative scene · Algorithm · Generation model

1 Introduction

The generation of stage scenes is a complex implementation process, which involves many aspects of research. At present, a lot of corresponding studies have been done on the technology of wenjing layout in China. For example, Li Jinghan in Harbin Institute of Technology focuses on the practical application of automatic placement of entities in 3D scenes and study the problem of the spatial relationship visualization based on natural language understanding. Also she realizes the automatic placement of entities in 3D space by two-step method [1, 2]. Another example is Wang Wei, who has conducted research on three areas in this field [3]: 1. Orientation word recognition based on pattern matching. 2. Target and landmark identification based on dependencies. 3. scene generation automatically based on Java3D. The main focus of these researches is

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 170–177, 2019.
https://doi.org/10.1007/978-3-030-04585-2_20

to use natural language to analyze the spatial dependence relationship between three-dimensional scene entities, and then use the spatial three-dimensional model placement technology to display the required scenes.

The main content of our research is not only the above method for the corresponding scene transformation, but we design and build a complete system approach from creative object generation, style construction to space combination to generate the stage scene we need. We also use GANs technology, style migration technology, and space combination algorithm.

2 Related Work

In this paper, we use GANs to build the creative stage model we need, style migration technology to generate the stage object style we need and natural language technology to analyze the stage creative points to get the projectile, boundary mark and orientation words. Based on this information, we first plan the order in which the stage objects are built. Then we select the required composition method and plan the scene in a certain area. Then construct a design experiment and gradually plan the detailed construction process of the scene.

2.1 Model Generation Based on 3DGANs Algorithm

As shown in Fig. 1 below, we apply the GANs algorithm to the generation of the required 3D object model (idea, object). In the actual model selection process, combined with brainwave technology analysis, if we do not have our satisfactory object model, we use 3DGANs technology based on the model library as a training sample [5] to generate our satisfactory object model, and use the GANs algorithm to solve the two major shortcomings of model library: 1. The number of models is insufficient; 2. The selected model does not meet our needs. The GANs algorithm is mainly composed of two parts: generator and discriminator [4]. The generator is used to generate the model and the discriminator is used to identify the generated model, and finally decide whether to deposit and enrich our corresponding model library. In this way, driven by the small model library, we gradually generate the large creative model library we need. Figure 2 shows the generation of a 3D model of the recurrent adversarial network construction generator, while discriminator is basically a mirror image of the generator. Equation 1 shows the main algorithm of GANs technology.

$$\min_D \max_G V(D,G) = E_{\chi \sim pdata(\chi)}[\log D(\chi)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (1)$$

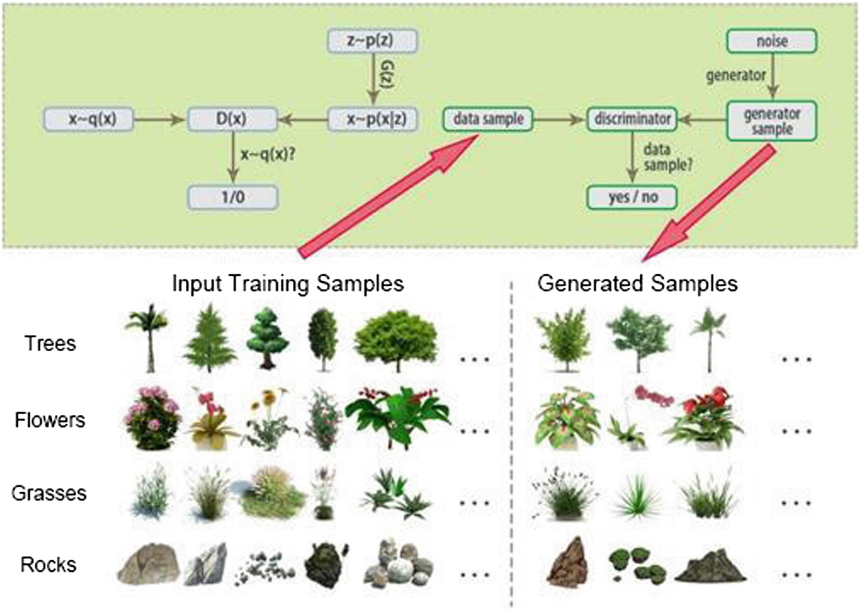


Fig. 1. 3DGANs algorithm for the generation model of 3D object

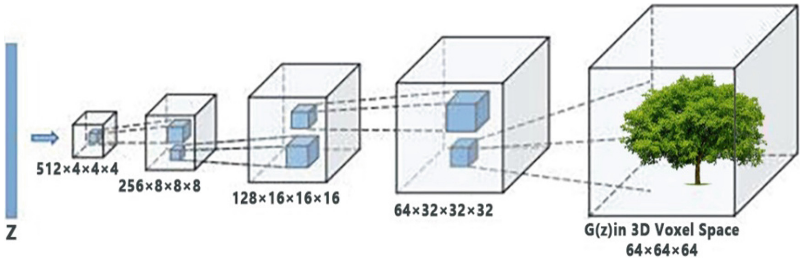


Fig. 2. 3DGANs generator model for cyclical network construction

In the process of using the GANs algorithm to generate the creative stage object, we consider the above-mentioned brainwave technology for creative evaluation (Fig. 5), which is described in detail in the following Sect. 2.4.

2.2 Generating Model Style Based on Style Transfer

After generating the required objects, we generate a model based on the style transfer to generate the required object map style. Style Transfer network facilitates the fusion of content and style, so that the generated images are added to a certain style in the case of guaranteed content [6, 7]. Figure 3 shows a map content target image and a texture style target image to be merged to generate the desired texture, so that when the content (such as texture) is unchanged, a certain style (such as: material, color, etc.) is introduced.

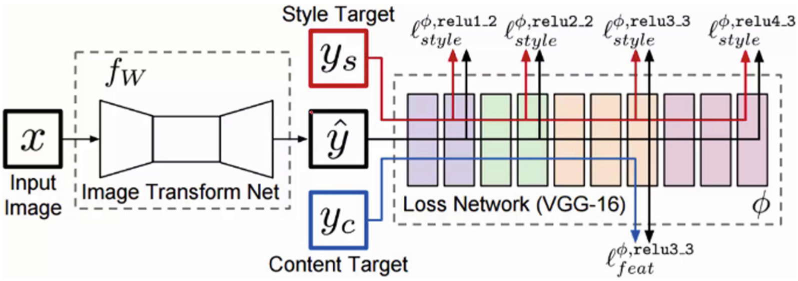


Fig. 3. Network of style transfer

2.3 Spatial Combination Generation Algorithm for Creative Scenes

When in the actual placement of specific objects, we semantically analyze the projectile, landmark, and orientation words to generate the creative scene object and adjust the proper position to complete the composition. If the landmarks and orientation words cannot be distinguished semantically, we set the landmark and orientation words of the corresponding objects using the random algorithm to complete the creative scene composition, according to the object composition order and composition method in the given creative works. The algorithm is shown in Fig. 4.

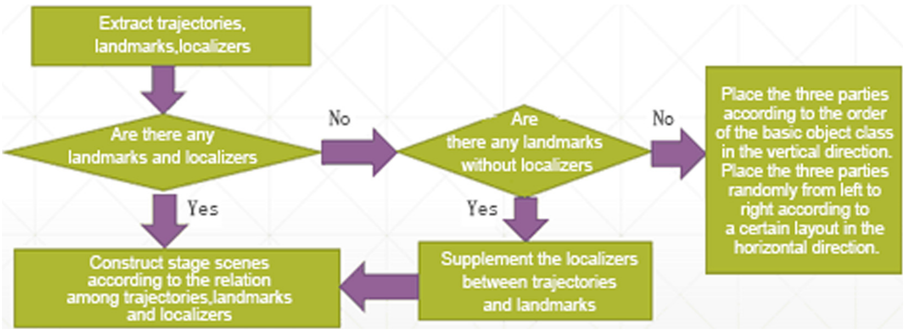


Fig. 4. The arranging algorithm for creative composition generation model

2.4 Evaluation of Creative Scenes

Among the many creative scenes generated, we should know what creative scenes we need. We use EEG to evaluate and select creative scenes. EEG is used to record changes in the electrical activity of the brain when it is an overall reflection of the electrophysiological activity of the brain's nerve cells on the surface of the cerebral cortex or scalp. [8, 9] Here, we mainly use EEG's P300 cognitive event-related potential combined with fast sequence visual presentation (RSVP) algorithm [10] and CNN convolutional neural network for classification training and analysis to extract the first sensory brain. The electrical event is the target feature [11], and the evaluation decision of the creative scenario we need is performed, as shown in Fig. 5.

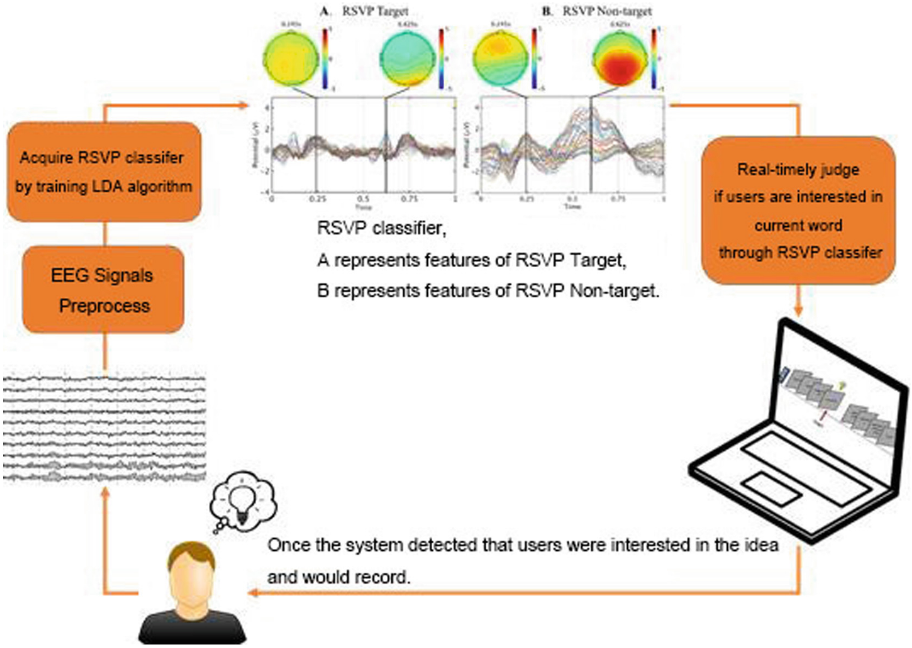


Fig. 5. Creative process evaluation based on RSVP algorithm

3 Experimental Results

3.1 Define the Order of the Objects in the Creative Work (Scene)

Building a scene in real life is also done by following a process from the bottom up. The construction of virtual creative scenes is also the same. The order of construction is to first lay the foundation, then cover the vegetation, then build the building, then put everyday tools and finally place the characters and animals. The basic placement sequence is shown in Fig. 6.

Following the above sequence, in the three-dimensional coordinate system of (x, y, z), the positional relationship of the model on the z-axis in the creative work of the simulation engine can be solved.

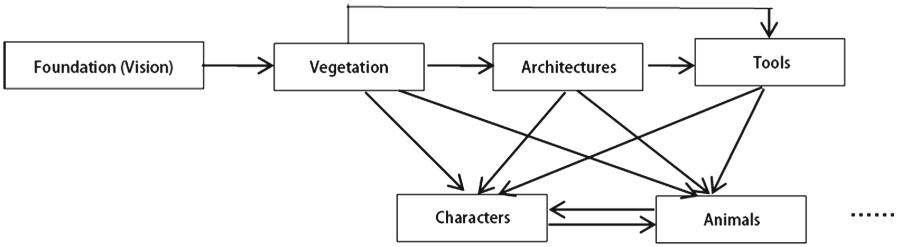


Fig. 6. The build order algorithm of the scene objects

3.2 Composition Method of Creative Work Construction

Basic composition method: pyramid composition method, S-shaped composition method, diagonal composition method, cross composition method, horizontal composition method, full composition method and symmetric composition method.

The construction of creative scenes is always arranged around a certain theme and meaning. For example, the creative scene of "Small Bridges and Waters", the bridge should be above the river and the characters should be above the bridge. These two arrangements follow the pyramid composition method (that is, the layout principle of a large model below the small model), and then the house follows the pyramid composition method alongside the river (the z-axis value of the house model is greater than that of the river) and horizontal composition (houses, bridges, characters not overlap in (x, y) coordinates). Such a layout is in line with the real-life position placement logic.

3.3 Script Code Modularity

Model Searcher
Function: Use the preset <model classification table> to traverse the existing imported model file name to record the existing model and count the number of models. At the same time, all recorded model positions are cleared to zero (position 0, 0, 0). At this point, a <model asset table> is generated.

Model Classifier
Function: Use the preset <model classification table> to compare with the name of the model in the <model asset table> generated by the model searcher. At this point, a <asset classification table> is generated.

Model Collision Detector
Function: Used to detect whether two models are interspersed or collided in the world of simulation engine creative works.

Composition Selector
Function: Call the composition method according to the actual creative scene, as follows: Pyramid composition method, S-shaped composition method, diagonal composition

method, cross composition method, horizontal composition method, full composition method, symmetric composition method…

Creative Scene Aligner

Function: According to the classification order of the construction principle (as shown in Fig. 6), the model of the corresponding class is selected from the <Production Classification Table>, and the model is arranged according to the construction principle and the call composition selector. Each model is set and recorded in the <model schedule> to count the models that have been laid out.

The digital work generation process is shown in Fig. 7. The creative scene generated by scene model based on the series of key algorithms is shown in Fig. 8.

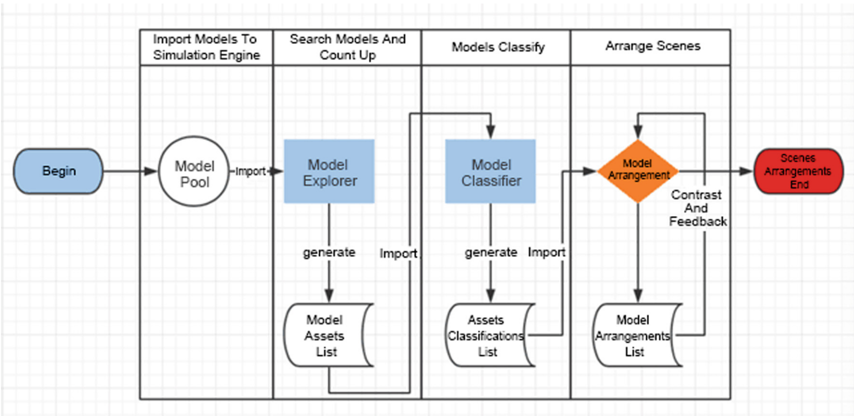


Fig. 7. Digital work generation process diagram

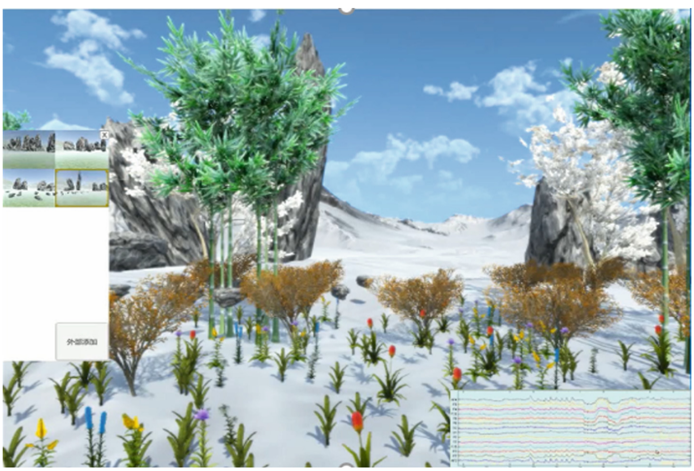


Fig. 8. Creative creation of scenes

4 Conclusion and Future Work

In this paper, we propose a generation model of the stage creative scene based on a series of algorithms. We hope to expand the model to enhance the intelligence of the stage digital simulation and digital performance in the future. On this basis, we will enhance the potential human-computer interaction ability of the stage by means of EEG technology, creative computing technology and neural network technology to improve the creativity of the stage, and make the generated digital stage more meet our actual needs.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204). We also acknowledge the solution from National Natural Science Foundation of China (61772254), Key Project of College Youth Natural Science Foundation of Fujian Province (JZ160467), Fujian Provincial Leading Project (2017H0030), Fuzhou Science and Technology Planning Project (2016-S-116), Program for New Century Excellent Talents in Fujian Province University (NCETFJ) and Program for Young Scholars in Minjiang University (Mjqn201601).

References

1. Li, H.: Research on spatial concept modeling based on natural language processing, pp. I–II. Harbin Institute of Technology, Harbin (2007)
2. Li, H., Li, S., Zhao, T., Han, Y., Ye, L., Li, L.: Study on automatic placement of entities based on natural language understanding. J. Electron. Inf. Technol. **29**(8) 2007
3. Wang, X.: Research on automatic placement technology of spatial entity described by natural language, p. I. Harbin Institute of Technology, Harbin (2014)
4. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
5. Wu, J., Zhang, C., Xue, T., et al.: Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling (2016)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)
7. Karayev, S., et al.: Recognizing image style. arXiv preprint arXiv:1311.3715 (2013)
8. Zhang, H., Wang, H.: Study on classification and recognition of multi-lead EEG signals. Comput. Eng. Appl. 24 (2008)
9. Keiper, A.: The age of neuroelectronics. The New Atlantis. Accessed 12 Feb 2016
10. Bigdely-Shamlo, N., Vankov, A., Ramirez, R.R., Makeig, S.: Brain activity-based image classification from rapid serial visual presentation. IEEE Trans. Neural Syst. Rehabil. Eng. **16**, 432–441 (2008)
11. Zhang, F., Mao, Z.-J., Huang, Y.-F., Lin, X., Ding, G.: Deep learning models for EEG-based rapid serial visual presentation event classification. J. Inf. Hiding Multimedia Sig. Process. **9**(1), 177–187 (2018)

On the Computability of Idea Creation of Creative Computing

Junjiang Lai[1] and Fuquan Zhang[2,3](✉)

[1] School of Mathematics and Data Sciences, Minjiang University, Fuzhou, China
439180193@qq.com

[2] School of Software, Beijing Institute of Technology, Beijing, China
8528750@qq.com

[3] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China

Abstract. Creative computing is an very important emerging technology in nowadays and computability is the necessary of researching creative computing. We demonstrate computability of idea creation of creative computing using the recursion function theory. The algorithm complexity for idea creation is also obtained.

Keywords: Idea creation · Creative computing · Computability
Computable function · Algorithm complexity

1 Introduction

As one of computation algorithms emerged, creative computing can be simulated the creative thinking of human being, it is extremely important in social life recently [1–4]. In creative computing creative idea creation is one of the kernel features. Zhang et al. proposed a method to generate creative ideas [5], there existed 5 steps in this method: (1) gathering relative data of motivation (2) clustering layer (3) idea generation layer (4) idea evolution layer (5) idea evaluation layer. And the experimental results show this method is good at building the creative idea which we need from natural language feature extraction.

In this paper, we investigate the computability of idea creation for creative computing [5]. The theory of computability mainly include decidability, computable functions and complexity. The problem is mainly to determine whether the equation has a solution; the computable function mainly discusses whether a function can be calculated, and establishes many mathematical models such as the original recursive function and the Turing machine to determine whether a function belongs to a computable function; the main problem of computational complexity is P = ? NP. computability theory tell us what can computers do in principle and what are their inherent theoretical limitations. Without loss of generality, we demonstrate the computability of idea creation theory which was put forward by Zhang et al. from the below three directions.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 178–183, 2019.
https://doi.org/10.1007/978-3-030-04585-2_21

In this paper, the following research is mainly carried out around the generation of computability of rough creative ideas, and the computable research is not carried out for the creative evaluation. The three aspects of research mainly include: (1) Based on motivation vocabulary and corpus segmentation, clustering generates the computability of the creative ideas ontology domain. (2) Improve the computability of the robustness of the creative ideas based on the Euclidean distance algorithm. (3) The computability of generating creative Ideas by extracting cluster members based on ontology domain and phrase and sentence structure.

2 Preliminary Knowledge

In this section, we give some preliminary knowledge for computability theory [6, 7].

Lemma 1. The following basic functions are computable:

(a) the successor function $s(x) = x + 1$,
(b) the constant function $C_n^{(n)}(x_1, \cdots, x_n) = m$,
(c) for each $n \geq 1$ and $1 \leq i \leq n$, the projection function U_i^n given by $U_i^n(x_1, x_2, \cdots, x_n) = x_i$.

From the above three kinds of basic functions we can obtain the recursion functions by compound method and recursion method. And the recursion functions can be expanded to general recursion functions(GRFs), which are computable functions. Here some common GRFs:

$$x + y; x \cdot y; x^y; |x - y|; \min(x, y); \max(x, y).$$

3 Main Results

3.1 Computability of the Ontology Domain of Creative Ideas Based on Motivational Vocabulary, Corpus Segmentation and Clustering

In the motivational layer of creative computing, we mainly use the acoustic principle to complete the input of speech motivation data to construct the list of motivational texts we need, and then collect relevant domain knowledge according to the list, and construct the relevant ontology domain. And our research on computability in this part is mainly focused on the latter.

We extract relevant corpora based on the motivation vocabulary list, or use the crawler algorithm to collect relevant corpus information to construct domain knowledge. Based on the domain knowledge, the K-means algorithm is used to cluster the motivational vocabulary list, and the ontology domain of the creative ideas is constructed to prepare for the later creative ideas generation. Below we conduct a computability argument for this part:

The principle of K-means algorithm is: set the sample set $D = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$, K-means algorithm minimizes the square $E = \sum_{i=1}^{k} \sum_{x \in C_i} ||\mathbf{x} - \mu_i||_2^2$ error, and obtains the cluster division $C = \{C_1, C_2, \cdots, C_k\}$, of which μ_i is the mean vector of the cluster C_i. This problem is an NP-hard problem [8]. Therefore, the iterative optimization is used to approximate the solution. The specific process is as follows:

(a) arbitrarily select k objects as the initial cluster center;
(b) separately calculating the center distance between each element in the data set and the selected cluster, and dividing the elements into corresponding clusters by the nearest neighbor principle;
(c) calculate the average of the objects in each cluster;
(d) repeating the above steps until the difference between the updated cluster center and the original cluster center is within a predetermined range or reaches a preset number of iterations;
(e) Output the centers of k clusters.

The complexity of the K-means algorithm is $O(NKT)$ [9], where N is the total number of elements, K is the number of cluster centers, and T is the number of iterations.

3.2 Computability of Robustness of Creative Ideas Based on Euclidean Distance Algorithm

The conception layer of creative computing mainly completes the research on the generation of creative ideas. Based on the motivational vocabulary list given in the previous stage and clustering to generate the ontology domain of the creative, we use the Euclidean distance to calculate the distance between the clusters. This makes it easy for us to select clusters with small distances when constructing creative ideas. The generated creative ideas are more relevant in the content of the composition, so as to avoid problems with grammar and semantics, and to improve the robustness of the creative ideas. Below we conduct a computability argument for this part:

Following the algorithm in [5], now we consider the computability for the step (3). Firstly define the Euclidian distance of two points

$$x = (\xi_1, \xi_2)^T, \quad y = (\eta_1, \eta_2)^T \in \mathbb{R}^2, \tag{1}$$

as

$$d(x, y) = \sqrt{(\xi_1 - \eta_1)^2 + (\xi_2 - \eta_2)^2}. \tag{2}$$

$d(x, y)$ is compounded by GRFs, it is GRF and then computable.

Secondly define the Hausdorff distance of two points sets

$$A = \{x_1, x_2, \cdots x_m\}, x_i \in \mathbb{R}^2, 1 \le i \le m,$$
$$B = \{y_1, y_2, \cdots y_n\}, y_i \in \mathbb{R}^2, 1 \le i \le n, \tag{3}$$

with

$$D(A,B) = \max\{h(A,B), h(B,A)\}, \tag{4}$$

where

$$\begin{aligned} h(A,B) &= \max_{1 \le i \le m} \min_{1 \le j \le n} d(x_i, y_j), \\ h(B,A) &= \max_{1 \le j \le n} \min_{1 \le i \le m} d(x_i, y_j). \end{aligned} \tag{5}$$

Obviously, $D(A,B)$ is a computable function.

Assume we have N sets denote as A_i, $1 \le i \le N$, denote $\ldots.k_i.\ldots$ as the number of points in A_i, x_j^i, $1 \le j \le k_i$ as elements in A_i, let $A_i = \bigcup_{j=1}^{k_i} \{x_j^i\}$. Based on the following 3 steps we can construct sentences with highly correlated semantics.

(a) Sort $A_i, 1 \le i \le N$ and define $D_{pq} = D(A_p, A_q), 1 \le p, q \le N, p \ne q$, once $D_{i_1 i_2} = \min\limits_{\substack{1 \le p,q \le N, \\ p \ne q}} D_{pq}, i_1 < i_2$, then A_{i_1} and A_{i_2} have the highest correlation.

Here, the algorithm complexity to obtain i_1, i_2 is $O(n^4)$ if we assume $N = k_i \equiv n$.

Let $\Lambda = \{1, 2, \cdots, N\}, G = \{i_1, i_2\}$, we can resort set $A_i, 1 \le i \le N$ from high correlation to low correlation as $A_{i_1}, A_{i_2}, A_{i_3} \cdots, A_{i_N}$ by following algorithm:

$$\begin{aligned} &For \quad m = 3 : N \\ &i_m = \arg \min_{i \in (\Lambda \backslash G)} \min_{j \in G} D(A_i, A_j) \\ &G = G \cup \{i_m\} \\ &End \end{aligned} \tag{6}$$

Noting that function $\arg(\cdot)$ is a GRF, so function i_m is computable. And the complexity for above algorithm is also $O(n^4)$.

(b) Select M points from sorted A_i and for any $x_{I_1}^{i_1} \in A_{i_1}, 1 \le I_1 \le k_{i_1}$, $x_{I_2}^{i_2} \in A_{i_2}$, $1 \le I_2 \le k_{i_2}$, we get $d_{I_1 I_2} = d(x_{I_1}^{i_1}, x_{I_2}^{i_2})$. Then sort $d_{I_1 I_2}$ as

$$d_{I_{1,1} I_{2,1}} \le d_{I_{1,2} I_{2,2}} \le \cdots \le d_{I_{1,M} I_{2,M}} \le \cdots, \tag{7}$$

define corresponding two sequences $W_{i_1} = (x^{i_1}_{I_{1,1}}, x^{i_1}_{I_{1,2}}, x^{i_1}_{I_{1,3}}, \cdots, x^{i_1}_{I_{1,M}})$, $W_{i_2} = (x^{i_2}_{I_{2,1}}, x^{i_2}_{I_{2,2}}, x^{i_2}_{I_{2,3}}, \cdots, x^{i_2}_{I_{2,M}})$, algorithm complexity for above sort is $O(n^2)$. Let $G = \{i_1, i_2\}$, we have

$$\begin{aligned}
&For \quad m = 3 : N \\
&n_* = \arg\min_{n \in G} D(A_n, A_{i_m}), \\
&\text{for every } x^{n_*}_{I_{,j}} \text{ in } W_{n_*}, 1 \leq j \leq M, \\
&\text{solve } x^{i_m}_{I_{m,j}} = \arg\min_{x \in A_{i_m}} d(x, x^{n_*}_{I_{,j}}), \\
&\text{let } W_{i_m} = (x^{i_m}_{I_{m,1}}, x^{i_m}_{I_{m,2}}, x^{i_m}_{I_{m,3}}, \cdots, x^{i_m}_{I_{m,M}}), \\
&G = G \cup \{i_m\}, \\
&End
\end{aligned} \tag{8}$$

Since $\min(\cdot)$ and $\arg(\cdot)$ both are GRF, functions n_* and $x^{i_m}_{I_{m,j}}$ are computable. We can conclude $O(n^4)$ algorithm complexity here. Thus for each set A_{i_n}, $1 \leq n \leq N$, we get a sub sequence of point denote with

$$\begin{aligned}
W_{i_1} &= (x^{i_1}_{I_{1,1}}, x^{i_1}_{I_{1,2}}, x^{i_1}_{I_{1,3}}, \cdots, x^{i_1}_{I_{1,M}}), \\
W_{i_2} &= (x^{i_2}_{I_{2,1}}, x^{i_2}_{I_{2,2}}, x^{i_2}_{I_{2,3}}, \cdots, x^{i_2}_{I_{2,M}}), \\
&\cdots \\
W_{i_N} &= (x^{i_N}_{I_{N,1}}, x^{i_N}_{I_{N,2}}, x^{i_N}_{I_{N,3}}, \cdots, x^{i_N}_{I_{N,M}}).
\end{aligned} \tag{9}$$

(c) Choose one element from W_{i_n}, $1 \leq n \leq N$ we get a new points set

$$x^{i_1}_{I_{1,j}}, x^{i_2}_{I_{2,j}}, x^{i_3}_{I_{3,j}}, \cdots, x^{i_N}_{I_{N,j}}, \quad 1 \leq j \leq M. \tag{10}$$

Given a fixed j, it can generate a sentence based on the sorted $A_{i_n}, 1 \leq n \leq N$, thus we have M sentences. Therefore, the algorithm for step 3) is computable and the total complexity is $O(n^4)$.

3.3 The Computability of Generating Creativity Ideas by Extracting Cluster Members Based on Ontology Domain and Phrase and Sentence Structure

On the basis of considering the Euclidean distance between clusters, the distance between the multi-cluster members is calculated by two pairs, and each pair of clusters takes the nearest N pairs of members. We are conceptually and conceptually, conceptually connected, and conceptually in the ontology. Depending on the nature of the attributes, we build our creative ideas based on the structure of the phrase or sentence. In this way, in the research of building creative ideas, we can further improve the accuracy of generating creative ideas. Because this part is a combination of phrases or sentences, it is naturally decidable to have clear implementation steps and methods within a limited time.

4 Conclusion and Future Directions

In this paper, the computability of idea creation of creative computing is established. Using the recursion function theory, we demonstrate the computable functions for algorithms of the parts of idea creation process. And the algorithm complexity is also concluded. All the algorithm steps of idea creation can be obtained similarly. Computability is the basis for creative computing. In the future, the computability on other aspects of creative computing will be considered by us.

Acknowledgements. The work of Junjiang Lai was supported by Natural Science Foundation of Fujian Provience, China (2016J01670). The work of Fuquan Zhang was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by the National Nature Science Foundation of China (Grant number: 61871204). We also acknowledge the solution from National Natural Science Foundation of China (61772254), Key Project of College Youth Natural Science Foundation of Fujian Province (JZ160467), Fujian Provincial Leading Project (2017H0030), Fuzhou Science and Technology Planning Project (2016-S-116), Program for New Century Excellent Talents in Fujian Province University (NCETFJ) and Program for Young Scholars in Minjiang University (Mjqn201601).

References

1. Jing, D., Yang, H., Tian, Y.: Abstraction based domain ontology extraction for idea creation. In: 2013 13th International Conference on Quality Software (2013)
2. Hugill, A.: Creative computing processes: musical composition. In: 2014 IEEE 8th International Symposium on Service Oriented System Engineering (2014)
3. Zhang, L., Yang, H.: Definition, research scope and challenges of creative computing. In: Proceedings of the 19th International Conference on Automation and Computing (2013)
4. Dean, D.L., Hender, J.M., Rodgers, T.L., Santanen, E.: Identifying good ideas: constructs and scales for idea evaluation. J. Assoc. Inf. Syst. **7**, 646–699 (2006)
5. Zhang, F., Mao, Z., Ding, G., Zhang, R., Lai, J.: Idea creation of creative computing based on embedding features and ontology. In: MISNC, SI, DS 2016, Union City, NJ, USA, 15–17 August 2016
6. Sipser, M.: Introduction to the Theory of Computation, 2nd edn. Thomson Course Technology (2006)
7. Cutland, N.: Computability: An Introduction to Recursive Function Theory. Cambridge University Press, Cambridge (1980)
8. Aloise, D., Deshpande, A., Hansen, P., Popat, P.: NP-hardness of Euclidean sum-of squares clustering. Mach. Learn. **75**, 245–248 (2009)
9. Blomer, J., Lammersen, C., Schmidt, M., Sohler, C.: Theoretical analysis of the k-Means algorithm–a survey. In: Algorithm Engineering. LNCS, vol. 9220. Springer, Cham (2016)

Preliminary Ideas for Intelligent Recognition Paths of Anomie Behaviors Among MOOC Learners

Shutang Liu[1(✉)], Fuquan Zhang[2,3(✉)], and Chunfen Jiang[4]

[1] Academic Affairs Office, Minjiang University, Fuzhou University Town, Fuzhou 350108, China
630670731@qq.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China
8528750@qq.com

[3] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

[4] School of Foreign Language, Minjiang University, Fuzhou University Town, Fuzhou 350108, China
376562437@qq.com

Abstract. This paper identifies the characteristics of anomie behaviors of MOOC learners through the cluster analysis of those anomie behaviors, and finds out the initial paths of intelligent recognition through online analysis of the background data characteristics of these anomie behaviors. The paper also suggests that these identification paths are not difficult to find, but there are still some profound reasons why many MOOC companies have not adopted those paths so far.

Keywords: Anomie behaviors of MOOC learners · Cheat
Intelligent recognition

1 Introduction

The study of learning anomie is based on anomie theory proposed by the American scholar Robert Merton. The core point of the theory is that it is the structure of society itself, rather than the free expression of human desires, that leads to individual behavior deviations. In recent years, various kinds of hack cheating software and delegate study online shops have emerged constantly in China. More and more types of cheating have arisen, such as multi-account cheating, identity authentication cheating. The author believes that based on the theory of anomie, these behaviors can be unified in accordance with the anomie of online course learning behavior, and the identification methods and post-processing methods of anomie behaviors are searched according to the theory of anomie.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 184–192, 2019.
https://doi.org/10.1007/978-3-030-04585-2_22

1.1 Background

With a series of documents issued by the Chinese Ministry of Education, including "Suggestions on Strengthening the Application and Management of Online Open Courses in Colleges and Universities" in 2015, "Suggestions on Promoting the Recognition and Conversion of Credits in Higher Education" in 2016, and "Regulations on the Administration of Students in Higher Education Institutions" in 2017 (Chinese Ministry of Education Order No. 41), most Chinese universities have already incorporated online learning courses into credit management. But due to the lack of supervision in all aspects and immaturity in the development of the entire concerned industry, students' anomie in online courses is common and widespread. For example, the hack cheating software (Fig. 1) and the delegate study on line shop (Fig. 2) are ubiquitous. Although most MOOC platform facilitators would complain about the delegate study on-line shop to the e-commerce platforms, every once in a while these shops will open again by changing names and images. In the three major MOOC platforms in the United States, with the promotion of curriculum certificate, micro-degrees and other projects, all platforms consider the learners' integrity of learning and intervened in all kinds of cheating.

Fig. 1. A hack cheating software for MOOC

1.2 Related Work

At present, research on this area focused mainly on two aspects. One, it is ensured that the learner and the receiver of the certificate must be the same person. This seems not to be a problem under the circumstance of organizable off-line examination and video monitored written examination and interview. However, this is not practical at all, yet still important, for those non-European and non-American countries and regions, which cannot afford to organize examination offline, or written examination and interview under video surveillance. For example, Serpil Kocdar and Cengiz Hakan Aydin, two

Fig. 2. Two delegate study on line shops for MOOC

scholars at Anadolu University in Turkey, discussed related issues in their paper “Quality Assurance and Accreditation of MOOCs: Current Issues and Future Trends”. The other aspect is the problem of technical cheating. Curtis G. Northcutt and Isaac L. Chuang of MIT, and Andrew D. Ho of Harvard University have done research on the detection and prevention of learners’ “multi-account cheating” (specifically refers to that fact that the learner registers multiple accounts and tries to test for correct answers through trial-and-error procedure).

Currently, there is practically no specific research in this area in China, but it has been mentioned in many studies on MOOC research. Cheng Yu, Han Yuanyuan, and Liu Jian have similar views on the reasons for the phenomenon of hack cheating in the online learning process of college students. They believe that there are two major reasons for this phenomenon: lack of interest and task requirements. Based on humanistic learning theory, social constructivism theory and situational cognitive learning theory, Yang Shan and Chen Huihui analyzed and summarized the factors affecting college students’ online learning, and finally proposed four coping strategies. As research on this topic is in its early stage, it would be worthwhile to do further and more detailed and profounder research as follows.

2 Major Anomie Behavior Clustering

The premise of identifying the anomie behavior of online course learners is to accurately cluster the anomie behaviors. Based on the previous studies, according to common features in the background data, the author deems that the following clustering can be included.

2.1 Anomie of Non-self Learning

This kind of anomie is the delegate study on line behavior mentioned above. This is completely a cheating behavior. The general characteristic of this behavior is that students, whose only purpose is to obtain the credits, do not conduct any online courses or take online exams. After the course is completed, there is no understanding of the course content and related knowledge. Some of these behaviors are achieved by purchasing from delegate study online shops. Some are simpler in the sense that a number of classmates selected a course at the same time, and then one student is arranged to watch the course videos on multiple computers for the rest classmates. If the course test is not conducted in a unified time, the same student is arranged to take all the tests.

2.2 Anomie of Watching Instructional Videos

There are mainly two types of such anomie. One is that the student's online course learning process is actually done by himself, but the learner is doing things unrelated to the course while the online course video is played. It is difficult to fully judge students' behaviors as cheating in this situation. From the practical point of view, some learners can complete most video study of courses, as well as online coursework, and conduct online courses test on their own. Only in the chapters that they are not interested in, there is such kind of anomie behavior. The other is to directly commission the delegate study online shop to play the teaching video or use the plug-in hack cheating software to cheat for the video watching data flow, but the videos are not actually played. This type of anomie is part of non-self learning as I mentioned in the Sect 2.1.

2.3 Anomie in Online Exams

In fact, in 2016, multi-account registration cheating method pointed out in the article by Curtis G. Northcutt of the MIT has existed in China's MOOC platform in recent years. However, in addition to the methods used by Chinese MOOC platform to allow different learners to do different tests, to make the questions in a random order, and to mark and give the answers after the test expires, other more effective methods have been adopted. That is, the after-class testing questions only appear after watching the requested course videos, and the final test questions will not appear until the learner has watched all or the required percentage of the course videos. This explains why Chinese cheaters have anomie in watching instructional videos. Currently the anomie in online exam is mainly achieved through the use the hacker software—modifying the background records of the video watching by the hacker software, so that cheaters can directly access the exam without watching the teaching videos. After that, some powerful hacking software can still get the answer data to help the cheaters to finish the exam.

3 Intelligent Recognition of Anomie

The intelligent recognition of anomie behavior means that the back-end server of the MOOC platform directly recognizes the three kinds of anomie behaviors mentioned above and directly processes them after identification.

3.1 Face Recognition, Fingerprint Recognition, Mobile Verification or Email for Authentication

Face recognition and fingerprint recognition may still be a problem a few years ago, but now it is no longer a technical problem. Many MOOC companies are trying to enable this feature, but how does this ensure that the face imported by the learner is his/her own? In fact, this is no longer a problem in China. Due to the popularity of online payment in China, it can be said that most Chinese students, including freshmen, have Alibaba's Alipay account, which has already started the face recognition function. In China, as long as MOOC companies choose to cooperate with Alibaba, they can get the facial feature identification information of the learner. Of course, this is not a problem in Europe and America. Apple Company had the fingerprint verification for some functions of the Iphone. The MOOC companies only need to cooperate with Apple Company. Generally speaking, MOOC companies lack the relevant technical ability and original information of customers, but these two companies have.

Of course, we do not rule out learners who do not use online payment accounts or do not use Iphone. This can be verified by sending a verification code to the user's mobile phone or email address. It is natural that few would like to lend their mobile phones to others for a long time or reveal their email password.

3.2 Identification of the Anomie Behavior of Watching Teaching Videos

For the recognition of the purchasing of delegate study service online, the simpler way is to use the method of IP address location. For example, a learner logs in to a learning account in one place and within ten minutes switches to another city one thousand miles away. Because there is no economic benefit in stealing a study account at all, it is basically for certain that the learner has entrusted the task of online learning to an institution in another city. This identifying method has been used by many Chinese MOOC companies. But, if the learner and the online delegate learning institution are in the same city, this requires further analysis. According to the author's research, the mechanism of manual delegate online learning institution often considers the economic benefits only. It is necessary to complete the video watching task and other learning tasks delivered by the client in a short period of time, so the time for playing the learning videos will be relatively concentrated. The time for ordinary learners to play learning videos and perform other learning tasks is more dispersed, which makes the corresponding learning time distribution curve take on a wavy look. The following are the two types of learning time distribution curves (Figs. 3 and 4):

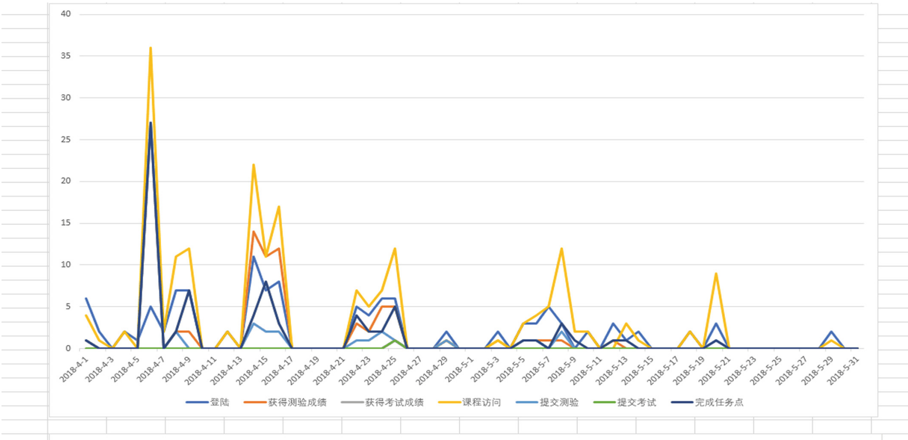


Fig. 3. An ordinary learner's learning time curve

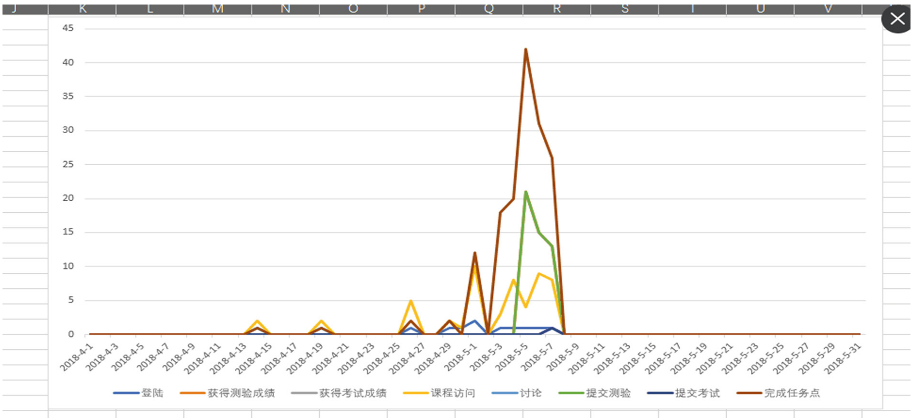


Fig. 4. A cheater's learning time curve

Admittedly, we can't directly judge the cheating merely because of the high concentration of learning time. Educators or teachers understand a very simple truth that time-intensive learning is not conducive and that enough time must be allowed for rest or thinking. Therefore, learners who have a high concentration of learning time must be reminded or identified again to prove that they are the learners who actually learn themselves.

Of course, it is very useful to do some basic settings to prevent the learner from being over-concentrated in learning time. For example, set the upper limit time for the video watching every day, or set some after-class questions after each video; before completing the tasks in the current section, the learner is not able to activate the next section. For those delegate learning institutions, it is impossible to complete the entrusted learning task in a short time, which increases their operating time costs. For ordinary learners, it can develop their good learning habits. If it is recognized that in a course there is a large amount of learners whose learning time is over-concentrated, it is also recommended that the responsible person in charge of the course arrange some additional control settings in some learning time and learning process.

3.3 Identification of Online Exam Cheating

Using the hacker software to bypass the learning of the course video is nothing more than a way to modify the data of the learner's video learning task, but it is not possible to falsify the basic data of the learner's knowledge point learning task. By just a simple comparison that could be found. It is not difficult to judge the cheating by using the hacker program to get the answer. That is, the exam questions or the after-class questions are completed too quickly. In real work, the author has found that some hacking software can even complete many multiple-choice questions in one second, and the answer is completely correct. Cheaters can be found by randomly detecting the interval between two successive exam questions, but the lower limit of the interval time must be set at first. The lower limit of the total answer time can also be used as a reference.

Identification of Social Anomie

As is mentioned earlier, there is no excessive technical barrier to the intelligent identification of online course learners' anomie behavior, and a more important factor is that MOOC companies are considering their operating costs. For instance, as time goes by, hackers always can find new vulnerabilities, strengthening MOOC platforms to prevent hack software definitely increases operating costs. MOOC was not originally initiated by nonprofit organizations. As long as an offline exam or online unified time exam (similar to the TOFEL) is conducted for those who want a micro-degree or course certificate, all technical means are not necessary.

However, many MOOC platforms have also found that, without updating the test database, learners get higher and higher scores, as time goes by. Because it is unlikely to prevent learners who have already achieved results from sharing the answers to a professional question and their personal opinions on various social media online. As time goes by, learners no longer have to register multiple user accounts to try the trial-and-error method; all they need to do is to search directly in the online search engine. If the learners' scores on a MOOC course no longer follow a normal distribution pattern, then it can be identified as a bad course. When most of the courses on a MOOC web are bad, the web site can be identified as a bad one.

Conclusions and Future Work

As is mentioned above, face recognition, fingerprint recognition, mobile verification or email verification can fully verify the identity of online learners. Through monitoring the location of the learner's IP address and detesting the concentration of learning time, it can be determined whether the learner is cheating in the video learning process. By comparing the learner's learning data before and after, it can be determined whether the learner has falsified the learning data. Through the learner's time interval and total time in answering questions in the test, it can be identified whether the exam taker cheats or not in the exam.

In general, because the research on related topics has just started, some specific intelligent recognition algorithms are not yet fully mature and need to be further explored. In addition, since there is no law in China to determine that "helping MOOC learners to cheat" is illegal, now formal MOOC companies are still helpless to "delegate online learning institutions". However, the social ethical changes which has brought about by the MOOC revolution still require further research and discussion.

Acknowledgments. The paper is supported by the foundation of Fujian Province Educational Science "Thirteenth Five-Year Plan" 2018 Project—"Research on the college students' anomie of online courses learning and intervention of their online courses learning" (No. FJJKCGZ18-850, Key funding project), Young and Middle-aged Teacher Educational and Scientific Research Project of Fujian Province—"Research on the college students' anomie of online courses learning and intervention of their online courses learning", and the Teaching Reform Research Project of Minjiang University in 2018—"The Interventional Teaching Reform aimed at the online courses learning anomie of college students" (No. MJU2018A005).Author ContributionsShutang Liu drafts the whole article, and accepts direct responsibility for the manuscript. Fuquan Zhang proposed the first step of the paper proposal, and contributes the conception of intelligent recognition, and gives the direction and suggestion for revising the article critically. Chunfen Jiang revises the translation of the paper, and is responsible for the language of the paper.

References

1. Kocdar, S., Aydin, C.H.: Quality assurance and accreditation of MOOCs: current issues and future trends. In: Proceedings of Open Education Global (2015)
2. Corrigan-Gibbs, H., Gupta, N., Northcutt, C., et al.: Deterring cheating in online environments. ACM Trans. Comput.-Hum. Interact. (TOCHI) **22**(6), 28 (2015)
3. Hart, M., Friesner, T.: Plagiarism and poor academic practice–a threat to the extension of e-learning in higher education? Electron. J. E-learn. **2**(1), 89–96 (2004)
4. Xiong, J., Wu, L.: Ethical analysis on application of massive open online courses (MOOC): concept, necessity and methodology. In: International Conference on Education, Management, Commerce and Society (EMCS-15), pp. 831–835. Atlantis Press (2015)
5. Chen, H.H.: Analysis of the Factors Affecting College Students' Online Course Learning Satisfaction. Normal University, Nanjing (2017)
6. Li, Q., Liu, N.: Research on MOOC quality assurance system. Open Educ. Res. **21**(05), 66–73 (2015)
7. Han, Y.Y.: Research on the status quo and problems of large-scale open online courses in the United States. Zhejiang Normal University (2015)
8. Yang, S.: Research on the promotion strategy of online course learning participation. Southwest University (2015)
9. Liu, J., Jiang, W., Zhang, M.Q., Li, X.X.: Study on the factors affecting college students' cheating in examinations. China Univ. Teach. (07), 85–88 (2010)
10. Liu, Y.: Analysis on the status quo of online course development and the influencing factors of teaching effect in China. China Educ. Technol. Equipment **02**, 116–118 (2018)

11. Liu, B., Zhang, W.L.: Research on the influencing factors and structure of online course learning experience. Mod. Educ. Technol. (09), 107–113 (2017)
12. Hui, Z.Y., Liang, Y., Li, S.J.: The obstacles for students to conduct online learning: factor analysis research. China Educ. Technol. **03**, 49–55 (2013)
13. Wang, K.H., Mondal, S.K., Chan, K., et al.: A review of contemporary e-voting: Requirements, technology, systems and usability. Data Sci. Pattern Recogn. **1**(1), 31–47 (2017)
14. Chen, C.M., Xu, L., Wu, T.Y., et al.: On the security of a chaotic maps-based three-party authenticated key agreement protocol. J. Netw. Intell. **1**(2), 61–66 (2016)

Fruit Detection Using Faster R-CNN Based on Deep Network

Linjuan Ma[1], Fuquan Zhang[2(✉)], and Lin Xu[3]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
923785608@qq.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, People's Republic of China
8528750@qq.com

[3] Key Laboratory of Nondestructive Testing, Fuqing Branch of Fujian Normal University, Fuzhou 350300, People's Republic of China
71471418@qq.com

Abstract. Fruit detection is of great significance in the agriculture. Recently, deep neural network has been widely studied in fruit detection. In our paper, we present a new approach of fruit detection, which uses Faster R-CNN based on deep network AlexNet. Our aim is to develop a fast, accurate and convenient model to detect and classify the fruit. Data augmentation is also used to increase the fruit dataset and avoid the overfitting in some degree. We use the fruits-360 dataset to perform the experiments.

Keywords: Fruit detection · Faster R-CNN · Deep neural network

1 Introduction

With the development and progress of artificial intelligence, fruit detection can be used in machine intelligence systems in the field of machine vision. Besides, fruit detection is of great significance in the agriculture, which can improve labor efficiency and market price competition. At present, there are variety of technologies to detect fruit. Such as Bulanon et al. propose to use visible and thermal images to fuse a new image to detect fruit [3]. And they also have proposed a new method of combination machine learning with a laser ranging sensor to detect fruit [4]. Also Kyosuke Yamamoto et al. [5] use image analysis and machine learning to detect tomato. And Lu et al. [6] use local binary pattern feature together with hierarchical contour analysis to detect immature citrus fruit. However, they mainly focuses on extracting fruit image feature combination with machine learning to detect fruits.

In our paper, deep convolutional neural network combined with faster r-cnn algorithms is applied to detect fruit. And we use fruits-360 dataset to perform the experiments, which contains 38409 images of fruits spread across 60 labels and 75 classes of the fruits.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 193–199, 2019.
https://doi.org/10.1007/978-3-030-04585-2_23

2 Related Work

Recently, deep network has been widely explored by many researchers in deep learning all over the world. Puttemans et al. [7] study a semi-supervised and fully automated fruit detector on account of boosted cascades.

In this section we will introduce the algorithm, deep network and technologies we have used in our work, including Faster R-CNN, deep network, AlexNet and data augmentation.

2.1 Faster R-CNN Algorithm

The Faster R-CNN algorithm is used in our work, as shown in Fig. 1 [1]. And there are two modules in it. Namely, Region Proposal Network (RPN) and classification model. When proposing regions, the RPN is used because of full-connected. However, in classification module, the individual regions is classified and the bounding box is regressed around the object.

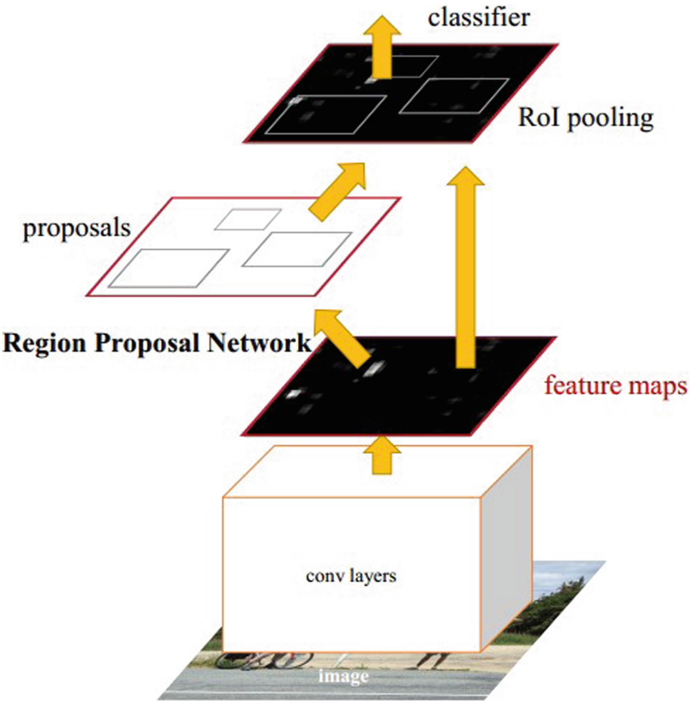


Fig. 1. An illustration of Faster R-CNN. (The figure is presented by Ren and readers can find the original figure in reference [1])

We input our image to pre-trained Convolutional Neural Network (CNN). The feature is extracted and the feature map is obtained. Then in the region proposal network (RPN), we slide small window over the convolutional feature map, which is outputted by the last conv layer. Therefore, region proposals are generated. And the mapped feature of the small window is fed into two layer. Namely, box-regression layer and box-classification layer, which are sibling FC layers [2]. The function of ROI pooling is to generate fixed-size anchor boxes. RPN is trained end-to-end by stochastic gradient descent (SGD). Also, RPN and R-CNN components share the convolutional layers. Besides, Non-maximum suppression is used to eliminate excess boxes in order to find the best object detection location.

2.2 Deep Network and Training

Recently, many researchers have applied ImageNet in pre-training CNN features and have obtained great progress in image processing, including image classification and image captioning [8].

However, we also need to fine-tune the network with our data. Only in this way can we take advantage of the features learned from the large-scale dataset.

As depicted in Fig. 2, our architecture of CNN we used is AlexNet eight learned layers contained. Also weights included. It is convolutional in the first five layers but fully-connected in the second three, which we call it the FC layer [2]. As Hinton et al. proposed, the last FC layer convey the output to a 1000-way softmax, which corresponds to the distribution of the 1000-class lables.

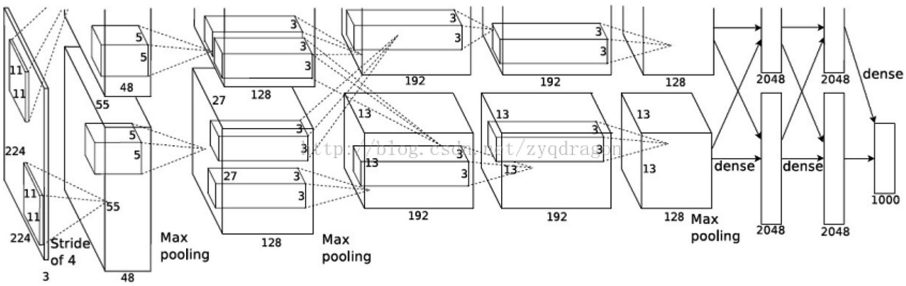


Fig. 2. An illustration of the structure of AlexNet.

The structure of AlexNet is made up of two parts because of two GPU training. One GPU runs at the top of the layer in the figure when the other [2] at the bottom of the layer. It is only in certain layers that they communicate with each other. The function of each layer is introduced in part 3 in detail. (The figure is presented by Krizhevsky et al. and readers can find the original figure in reference [2]).

2.3 Data Augmentation

It is very common to extend the variability of the training data using data augmentation. We can manually use label-preserving transformations, which the dataset is enlarged. Data augmentation can also reduce overfitting on models, where we increase amount of training data using information only in our training data. The traditional ways of augmenting image data is to perform geometric and color augmentations, including horizontal-vertical flipping, image random scale, random crop, PCA jittering and changes to image color. In our work, the PCA augmentation technique is used, which presented in AlexNet [2]. Firstly, according to the three colors of RGB, the mean and standard deviation is calculated. Then, the covariance matrix is calculated on the whole training set. The feature vector and feature value is obtained when we conduct feature decomposition. And it is used for PCA Jittering.

3 Experiments on Fruits-360

The architecture of our fruit detection and the results of end-to-end performance of fruit detection is introduced in this part.

Our dataset is based on fruits-360, a dataset of images containing fruits. The dataset has 38409 images of fruits spread across 60 labels. There are 75 classes of the fruits. And the framework in our work is explained in Fig. 3. Firstly, we input an image of 3-channel. Then it is conveyed to a series of convolutional layers. Therefore, it can propose region of interest boxes. And then each box [2] is conveyed through FC layers. Class probability together with a finer bounding box [1] around individual objects is returned in this layers. During training, the input image of green boxes is ground truth, which is used in the RPN and the R-CNN layers. And the output applies specific detection threshold for a class. However, non-maximum suppression is used to eliminate the results overlapped.

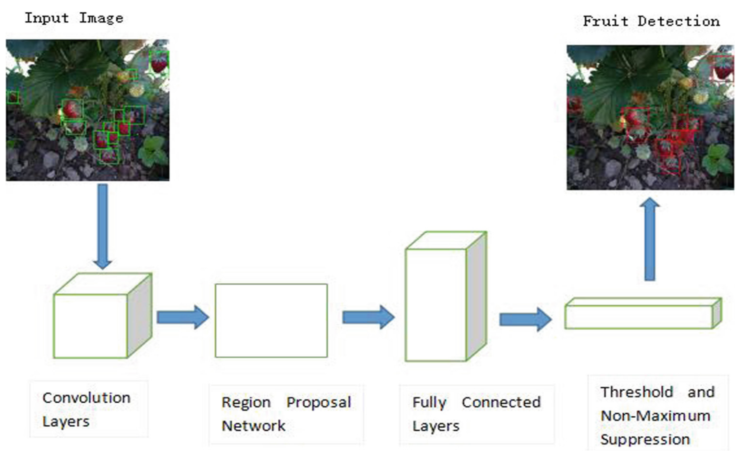


Fig. 3. The framework of our fruit detection

The size of input image is resized 227 * 227 to the AlexNet. And each layer of our AlexNet is illustrated in Table 1. When we training AlexNet network, ReLU and Norm are added after each conv layer. ReLU is used as the activation function of CNN. Its effect exceeds Sigmoid in the deeper network. Also the training speed is faster because of gradient descent. Dropout is used after the last two full connection, which can randomly delete some hidden neurons in the network to avoid over-fitting. At the same time, a competitive mechanism is created by the Local Response Normalization LRN) layer for the local neurons activity to enhance the generalization ability of the model.

Table 1. Layers of AlexNet

Layer type	Dimensions	Outputs
Conv1	55 × 55 × 96	96
ReLU		
Pool1	27 × 27 × 96	
Norm1		
Conv2	27 × 27 × 256	256
ReLU2		
Pool2	13 × 13 × 256	
Norm2		
Conv3	13 × 13 × 384	384
ReLU3		
Conv4	13 × 13 × 384	384
ReLU4		
Conv5	13 × 13 × 384	256
ReLU5		
Pool5	6 × 6 × 256	
Full6	4096 × 1	4096
Relu6		
Drop6		
Full7	4096 × 1	4096
ReLU7		
Drop7		
Full8	1000 × 1	1000

The sample images of ground-truth annotation are green bounding boxes as shown in Fig. 4.

Fig. 4. Sample images of ground-truth annotation on the fruits-360 dataset

And the results of end-to-end performance of our fruit detection are as shown in Fig. 5, while red bounding boxes are detection results.

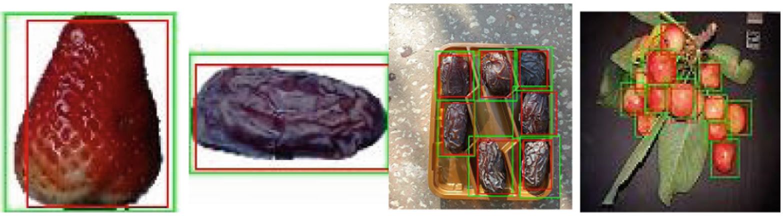


Fig. 5. Results of fruit detection on the fruits-360 dataset

4 Conclusion

We introduce a fruit detector based on fruits-360 dataset applying Faster R-CNN in our paper, which is the state-of-the art detection framework. Fruit detection is of great significance in the agriculture, which can improve labor efficiency and market price competition. In the future, we will improve the performance of our deep network to detect more fruits and meet the needs in our actual life.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204). We also acknowledge the solution from National Natural Science Foundation of China (61772254), Key Project of College Youth Natural Science Foundation of Fujian Province (JZ160467), Fujian Provincial Leading Project (2017H0030), Fuzhou Science and Technology Planning Project (2016-S-116), Program for New Century Excellent Talents in Fujian Province University (NCETFJ) and Program for Young Scholars in Minjiang University (Mjqn201601).

References

1. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster RCNN: towards real-time object detection with region proposal networks. CVPR, arXiv:1506.01497v3 (2016)
2. Krizhevskym, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
3. Bulanon, D.M., Burks, T.F., Alchanatis, V.: Image fusion of visible and thermal images for fruit detection. Biosyst. Eng. **103**, 12–22 (2009)
4. Bulanon, D.M., Kataoka, T.: Fruit detection system and an end effector for robotic harvesting of Fuji apples. Agric. Eng. Int. CIGR J. **12**(1), 203–210 (2010)
5. Guo, W.: On plant detection of intact tomato fruits using image analysis and machine learning methods. Sensors **14**(7), 12191–12206 (2014)

6. Lu, J., Lee, W.S., Gan, H., Hu, X.: Immature citrus fruit detection based on local binary pattern feature and hierarchical contour analysis. Biosyst. Eng. **171**, 78–90 (2018)
7. Puttemans, S., Vanbrabant, Y., Tits, L., Goedemé, T.: Automated visual fruit detection for harvest estimation and robotic harvesting. In: IPTA (2016). https://doi.org/10.1109/ipta.2016.7820996
8. Cireşan, D., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. Arxiv preprint arXiv:1202.2745 (2012)

Using Quantum Thinking to Construct a Musical Appreciation System

Mei Peng[1], Ding Gangyi[1], Jin Qiankun[1], and Zhang Fuquan[1,2](✉)

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
592126776@qq.com, 1961536122@qq.com, 8528750@qq.com, dgy@bit.edu.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, People's Republic of China

Abstract. With the richness of material and cultural life, participating in music activities and enjoying music works has become an indispensable activity in people's leisure time. The rich emotions in music have different performances according to individual differences. After the composers create the music, the performer's deductive processing and then the emotion is sublimated by the viewers finally. This paper proposes a new emotional system for basic emotional joy, anger, sadness, joy, etc., and interprets the superposition state of emotional expression in detail. The paper believes that the generation of consciousness and the resonance of thinking are precisely the result of the quantum collapse of mutual entanglement.

Keywords: Music appreciation · Emotional system · Quantum consciousness Emotional evaluation

1 Introduction

1.1 Existing Major Music Appreciation System

In music appreciation, the emotional system of music is generated through the interaction of three stages of participants. Composers give the initial emotion of music through personal life experience, creative environment, and imagination during inspiration [1], etc. Performers produce a certain empathy with the composer in the process of expressing music. Because of the differences of the social consciousness, education level, professional characteristics and personality, appreciators have different resonances of a certain part of music when they watch the show of performances [2].

There is a certain correspondence between the structural level of music works and the psychological elements of music appreciation, as shown in Fig. 1 below. The psychological elements of music appreciation are divided into acoustic perception, emotional experience, imaginative association and understanding [3]. The structural level of musical works includes the surface layer and the inner layer.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 200–207, 2019.
https://doi.org/10.1007/978-3-030-04585-2_24

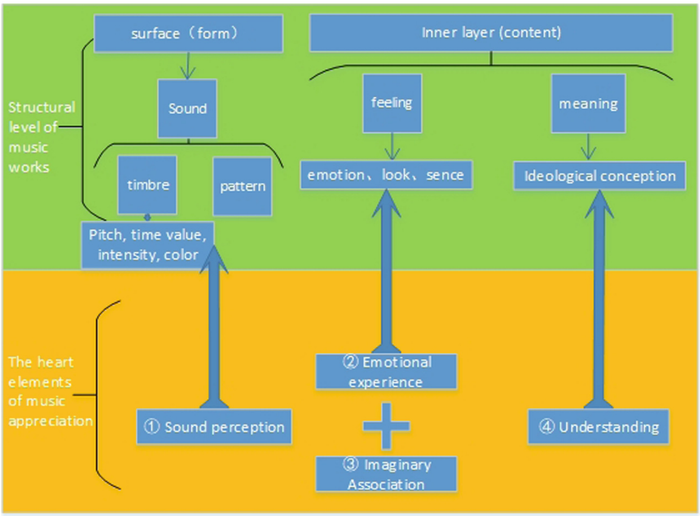


Fig. 1. Music appreciation traditional emotional system

1.2 The Relationship of Quantum Consciousness and Emotion

In the 1950s, the concept of quantum consciousness was mentioned in David Bohm's book Quantum Theory, which became a classic textbook in the field. Bohm uses the experience of listening to music to discuss the nature of consciousness. He believes that the feeling of movement and change constitutes people's musical experience. Bohm believes that this is the emergence of consciousness from the implicit order.

Quantum consciousness theory believes that classical mechanics cannot fully explain consciousness, and consciousness is a phenomenon of quantum mechanics, such as quantum entanglement and superposition [4]. There are a large number of electrons in the brain that are in quantum entanglement, and consciousness is generated from the periodic collapse of the wave functions of these electrons [5].

Penrose believes that the fundamental difference between the human brain and the computer may be caused by the uncertainty of quantum mechanics and the chaotic effect of complex nonlinear systems [6, 7]. The human brain contains a non-deterministic and naturally formed neural network system, which has the "intuition" that the computer does not have. It is the "fuzzy" processing power and the extremely high performance of this system. The traditional Turing machine is a deterministic serial processing system, although it can simulate such a "fuzzy" process, but the efficiency is too low [8]. The quantum computer and computer neural network systems under study are really promising to solve such problems and reach the human brain.

1.3 Simple Explanation of Emotions by Neural Networks

In many machine learning models, we tend to see an expression like “wx + b”. “x” represents a sample, and a sample typically has multiple attributes, so x is generally a vector. Geometrically, “wx + b” can be thought of as a hyperplane. From the algebraic point of view, it is a “weighted + biased” operation, “w” is the weight, and “b” is the bias [8]. We use this as an entry point, aiming at different emotions to superimpose different emotional factors on its dominant emotions. The goal is measuring the emotional expression of the influence of music expression and forming a new music emotion evaluation system.

2 Main Work

2.1 Participant Entanglement Definition

In the music appreciation system, there are three types of participants: composers, performers and appreciation [10]. In the emotional system, the basis for emotional resonance is the matching of social environment and class status. The music evaluation system is divided into four stages: music creation, music performance, music appreciation and music review [11]. Among the three types of participants, they interact with each other to form a closed loop of emotional influence. They present entangled states of each other and each role will affect the whole process of creation [12]. The relationship of performance, appreciation and commentary of the entire music is shown in Fig. 2.

In the process of creating music, the composer firstly generates certain ideological cognition and emotional sources. At some point, the composer’s inspiration bursts into a rich emotion that blends into the music work [13]. Some of these emotions are conscious and some are subconscious. Therefore, the emotions of musical works are more abundant and require individuals to comprehend [13].

After the composer has laid the emotional tone of the musical composition, it basically determines the performance form of the performers. For a complete musical work, the subtle fluctuations of emotions and continuous changes in details can only rely on personal comprehension. Sometimes, they just judge intuitively or react instinctively. It is an unconscious state. In the process of performance by performers, emotional empathy is related to the composer and can be seen as the formation of entangled states.

Appreciators watch the performances of the performers. Due to the individual differences between the viewers and the performers, different emotions will be formed. At this time, the appreciators will form a psychological reaction between the good and the evil and make their own judgments on the musical works to form knowledge. And this process is another form of entanglement.

The reaction of the appreciators is the formation of feedback, which will affect the re-creation and continued creation of the composers. As shown in Fig. 2, this has formed a closed loop of emotional transmission between the three participants.

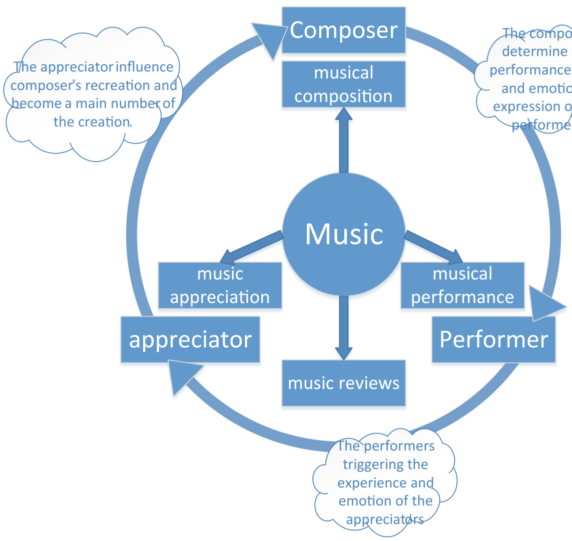


Fig. 2. Closed loop of emotional transmission

In the same way, the composers' creative conditions, performance and other factors will be used as a reference for the appreciation of the music works by the appreciators. The responders' reaction will also provide adjustment guidance for the performers' performance on the spot. Similarly, the performers' performance mode and emotional rendering also provide a basis for the composers to create musical works more expressive. Therefore, a reverse closed loop of emotional feedback between the three participants can also be formed.

2.2 Analysis of Consciousness and Expression of Emotion

Consciousness is the process of collapsing coherent electrons and reconstructing coherent electrons for new feelings [15, 16]. This is the interpretation of consciousness by the principles of quantum mechanics and the basis of quantum consciousness [17–19]. This article will explain the quantum phenomena produced by the three types of character emotions during the music appreciation process. This section will explain each of the issues mentioned in the previous section, such as the psychological emergence of the composers, the intuitive judgment of the performers and the instinctive reaction, which is the stat e of unconsciousness, and the formation of the viewers' psychology of the music–like or dislikes.

In the early stage of music creation, composers will have a concentrated psychological emergence in a certain period of time due to the influence of the creative environment and the limitations of life experience. It may be sadness or joy, love or disgust. It may be to create an atmosphere or image which is positive or negative. The composers' psychological emergence has formed a complex superposition state and there is a certain gradual or progressive relationship between emotions. The tone of the beginning of music works will lay the foundation for the development of subsequent emotions. This phenomenon is explained by emotional entanglement. In this way, the emotional elements will form a complex map on the musical works.

In the process of the music works being presented by the performers, the performers have certain differences with the music environment and life experiences of the composers, so the emotional understanding of the music works cannot be fully grasped. When they are expressing music, some parts rely on intuition and instinct to play according to their performance environment, personal psychological state, etc. This is an unconscious operation. The composers and performers with entanglement relationship have collapsed at this time, spanning the dimension of time and space. Appreciators will form their own reactions in the process of enjoying music according to the expression of the performers' musical works. Finally, the impression of the music works will be formed simply–"good and evil" reaction. Each viewer is a basic quantum unit in a superimposed state of "good and evil." If we interview enough audiences, we can summarize the advantages and disadvantages of the overall effect of the music works, although the opinions and expressions of each audience are different.

2.3 Analysis of Emotional Superposition State

The superposition of emotions can be said to be a generalized superposition. The rich emotions of human beings also make the superposition state more abundant. The combination of happiness and sorrows also makes the evaluation system of musical emotions more practical. If you simply express the four emotions of pleasure, anger, sorrow and joy, as shown in Fig. 3.

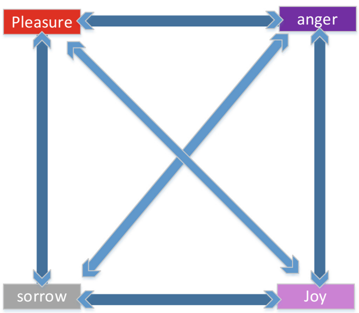


Fig. 3. The simple relationship between pleasure, anger, sorrow and joy

In the analysis of human emotions, this paper believes that all kinds of emotions coexist, but at a certain moment, an emotion occupies an active position. Based on this, we express the emotional component map of the joy of emotion as follows, as shown in Fig. 4.

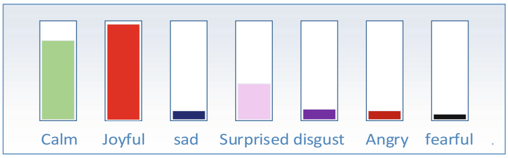


Fig. 4. Emotional composition of joy

As can be seen from Fig. 4, when happy, the happy mood dominates, and the other emotions are relatively weak, but not non-existent. Based on this theory, the emotional transformation can be explained to a certain extent. The generation of emotion is not a process from nothing, but a process of continuous change from existing emotions. From the perspective of the neural network, an emotional sample "a" generates an offset "b" because of the coexistence effects of other emotions in the case of weighting, which provides an effective guarantee for us to determine the method of emotional evaluation. The measure of emotion provides a theoretical basis for the study of the control of weak emotions' change in music appreciation.

2.4 Summary

This paper is important to establish an emotional evaluation system for music appreciation. The core components of the system are composers, performers and appreciation of the entanglement in the process of music creation, performance, appreciation and evaluation. Each character is indispensable and will influence on emotions of the music works. Secondly, the inspiration of the composers, the intuition of the performers' understanding of the music works and the unresponsiveness of the audiences' psychology are the process of quantum reaction. The different processes of consciousness formation also confirm that the quantum can make the expressiveness of music emotions richer and more attractive. The expression of emotions is only that some emotions occupy a dominant position. Other emotions do not play a major role, but they still exist, and can explain the changes between emotions. These studies can provide a basis for the emotional evaluation of music appreciation to a certain extent.

3 Experimental Concept

This part introduces our experimental conception of the above theory. The experiment is not perfect and there is still some work to be done.

First of all, the experimental background is based on the seven emotions of Fig. 4, which divides emotion into positive emotions and negative emotions. Therefore, happy, amazed, and calm are a group, otherwise, sad, disgusted, angry and fearful are another one. In each group, the emotions are superimposed and there is entanglement. Then, we will perform brain-energy capture on a large number of different individuals, including composers, performers and appreciators. Meanwhile we count a large number of targeted data. On the one hand, they reflect the whole process of the same track and unified tone emotions; on the other hand, they are emotional perceptions of different musical works and varied music environments, in order to determine the emotional waveform of the seven emotions. Next, we will build two neural networks for machine learning, used to continue to collect large amounts of EEG data and to classify and map music and emotions. Finally, on the basis of collecting and processing a large amount of data, the orthogonal map of the new music in each emotion is obtained. Then the data reference of the music appreciation evaluation system is provided according to the data trend and the peak size.

4 Evaluation of Results

1. Find existing EEG emotion data and process it as input data. After that, we will compare the training results with the results of the existing experimental capture data training to find the difference in the emotional label labeling.
2. Follow-up research work is also necessary, which is around the social value of this music appreciation emotional evaluation system to evaluate its practical guiding significance.

– Whether the quality of musical works will increase in future creations.
– Whether this assessment method has a positive social effect and has a positive impact on the whole society.
– Whether this method has promoted the status of Chinese music in the world.

Acknowledgements. This study was supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204).

References

1. Wang, X., Yang, C.: On the emotional factors in music performance art. Musical Concept (12) (2011)
2. Zhang, M.: On the emotional expression in the art of vocal performance. Brand (Lower Half) (3) (2015)
3. Feng, B.: Analysis of the expression art of emotion in music. Popular Lit. (2), 134 (2013)
4. Ding, G., Jin, Q., Mei, P., Zhang, F.: The opportunities and challenges of quantum computing. Biomed. J. Sci. Tech. Res. **6**(3), 1–3 (2018)
5. Kokosar, J.: Arguments and model for quantum consciousness, modification of quantum collapse, and panpsychism (2013)
6. Sbitnev, V.I.: Quantum consciousness in warm, wet and noisy brain. Mod. Phys. Lett. B **30** (28), 1650329 (2016)
7. Myroniv, B., Wu, C.-W., Ren, Y., Christian, A.B., Bajo, E., Tseng, Y.-C.: Analyzing user emotions via physiology signals. Data Sci. Pattern Recogn. **1**(2), 11–25 (2017)
8. Zuo, C.: Penrose's exploration of quantum mechanics of consciousness and its spiritual philosophical significance. Central China Normal University (2015)

9. Xiaofeng, W., Guohua, Y., Peng, L., et al.: A neural network music sentiment classifier construction algorithm based on relevance feedback. J. Northwest. Univ. (Nat. Sci.) **42**(1), 30–35 (2012)
10. Sloboda, J.A.: Music structure and emotional response: Some empirical findings. Psychol. Music **19**(2), 110–120 (1991)
11. Wiese, M.A., Holland, E.D., Barbee, E.R.: Musical system: US, US 8008560 B2 (2011)
12. Xie, X.: Research on the audience of dance creation to the market. China Academy of Art (2015)
13. Bella, S.D., Peretz, I., Rousseau, L., et al.: Development of the happy-sad distinction in music appreciation. Ann. NY Acad. Sci. **930**(1), 436–438 (2010)
14. Livingstone, S., Muhlberger, R., Brown, A., et al.: Changing musical emotion: a computational rule system for modifying score and performance. Comput. Music J. **34**(1), 41–64 (2010)
15. Conte, E.: Testing quantum consciousness. Neuroquantology **6**(2), 126–139 (2008)
16. Goradia, S.: Quantum consciousness the road to reality (6), 269–272 (2016)
17. Stenger, V.J.: The myth of quantum consciousness. Humanist **53**, 13–15 (1992)
18. Zheng, R.S.: Quantum consciousness - a new understanding originating from quantum psychology. J. Xuzhou Normal Univ. (2008)
19. Husain, S.: Universal field of quantum consciousness: an ontological nexus of classical and quantum physics. Neuroquantology **8**(3), 279–286 (2010)

VRAS: A Virtual Rehearsal Assistant System for Live Performance

Yufeng Wu[1], Gangyi Ding[1(✉)], Hongsong Li[1], Tong Xue[1], Di Jiao[1], Tianyu Huang[1], Longfei Zhang[1], Fuquan Zhang[1,2], and Lin Xu[3]

[1] Digital Performance and Simulation Key Laboratory, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
{wuyufeng, dgy, lihongsong, 2120161107, 2120161105, huangtianyu, longfeizhang, zfq}@bit.edu.cn
[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China
[3] Key Laboratory of Nondestructive Testing, Fuqing Branch of Fujian Normal University, Fuzhou 350300, People's Republic of China
71471418@qq.com

Abstract. The VR technologies are widely adopted for training purposes by providing the users with educational virtual experience. In this work, we propose an immersive VR system that help the choreographers and the dancers to facilitate their dance rehearsal experience. The system integrates motion capture devices and head-mounted displays (HMDs). The motions of the dancers, their partners, and the choreographers are captured and projected into a virtual dancing scene in an interactive frame rate. The dancers who are wearing the HMDs, are allowed to observe the synthesized virtual performances within a virtual stage space from several selected third-person views. These are the audience's view, the dancing partner's view, and the choreographer's view. Such synthesized external self-images augment dancer's perception of their dance performance and their understanding of the choreography. Feedbacks from the participants indicate the proposed system is effective and the preliminary experimental results agree with our observations.

Keywords: Dance rehearsal · Choreography · Immersive virtual reality
Motion capture

1 Introduction

The main task of dance performance is to induce aesthetic experience of the audiences. To achieve this goal, the choreographers and the dancers need to integrate many elements, including the dancer's movements, interactions between dancers, narrative, music, lighting, and costumes. Although dancer training systems are proposed to provide better training to the dancers, the modern technologies provide little help to this wearisome rehearsal process.

Our research work started from the interviews with the dancing teachers and students of Beijing Dance Academy. The dancing teachers have rich experience described the existing problems using traditional method in their daily training and the rehearsal

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 208–216, 2019.
https://doi.org/10.1007/978-3-030-04585-2_25

processes. It is a challenge to discover the problem by observing the entire dance through the video especially from one viewing angle, and it is a time-consuming process for the dancers or the choreographers to verify the formation of the dancers and the synchronization of the movements. The dancers would love to evaluate the dancer's performance by navigating in the dance scenes during the rehearsal process.

To address this issue, we present an immersive virtual reality system with three selected third-person views: the audience's view (AV), the partner's view (PV) and the choreographer's view (CV) according to the role elements (choreographer, dancer, and audience) involved in the performance. Han present augmented reality system for guiding arm movement in the first-person perspective and the vision is limited [9]. However, he proposed a third-person view tool by using a drone to track the users [10]. It has limitations on the range that the observable to the user.

The purpose of our system is to help the dancers not only integrate their own dance movement with the other performance elements but also provide several unprecedented perspectives to understand the choreography and evaluate the dancer's performance more efficiently and thoroughly. The system overview is illustrated in Fig. 1. Note that the dance rehearsal is often a highly iterative design process. The choreography would be revised constantly so that both the dancers and the choreographer should be able to communicate comfortably in this system.

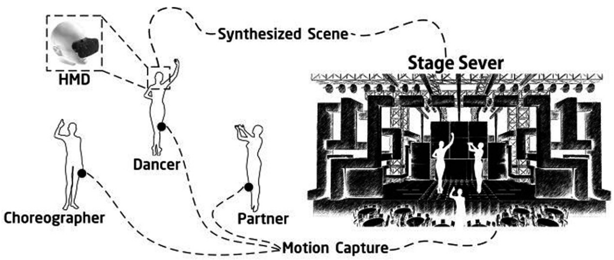


Fig. 1. System configuration. The dancers wearing the HMDs are allowed to observe the synthesized virtual performances within a virtual stage space from several selected third-person views, which belongs to the audiences (we use a virtual camera to simulate the audience's view because the audience is not moving), the dancing partners, or the choreographers.

2 Related Work

Applying VR/AR/MR technologies for dancer training has been widely reported in recent years. Some dance training systems collect dancer's motion data and compare the data with certain motion template in order to evaluate the dancer's performance [20]. In principle, such systems can help a learner (usually a non-professional learner) to master certain category of dance [8]. Kyan et al. presented a cave virtual reality system for ballet dance training, which compare the student's motion with the teacher's [12]. Similar approach is adopted by Chan et al. to guide students to improve their skills [4, 5]. Some dance training systems help the dancers by augmenting their perception. This approach is to provide a better tool than the mirror a dancer often used to observe his/her own movement [3, 14]. OutsideMe augmented external self-image by display on Oculus Rift HMD the dancer's self-image captured by a wide-angle camera [19]. The same method is used to design a training systems for ballet [7, 18]. And also, some research presented a mobile robot system for easy dance training [15] and tracking the human body with a real-time technology [1].

The above-mentioned training systems focus on single student training tasks. Other researcher pointed out that interactions among the dancers also require extensive training. Ho et al. presented a framework to generate virtual characters dancing with real dancers [6]. Their method provides the user with real-time visual feedbacks for interactive partner motion synthesis. Uejou et al. are also try to achieve this goal for ballroom dance training [13]. This is a good start to help the dancers improve their interactions with their partners. However, how to apply the modern technologies to improve the interactions among the real dancers are still unknown.

In this work, we assume that the dancers are well-trained to master the dance movement before the rehearsal, because of the main task of the virtual rehearsal system is to make a good dance rehearsal, instead of training a person without dance foundation to learn how to dance. The main challenge of rehearsal is to integrate all the performance elements according to choreographer's design. More importantly, the motions of the dancer, their partner, and the choreographers are all captured and integrated into the synthesized scene.

3 System Design

3.1 System Overview

To address to the dancer's needs during rehearsal process, we propose a prototype rehearsal assistant system that consists of a HMD (HTC Vive), three motion capture devices (Perception Legacy), and a virtual reality system build with a 3D game engine (Unity5).

The "Perception Legacy" motion capture device includes 17 inertial sensor nodes and the whole wearing weight is less than 230 g is illustrated in Fig. 2(a). Especially, it battery can last for four hours and the range of motion is 50 m. From above, it satisfied with dance training. The synthesized virtual scene is outputted to a HTC Vive HMD. It brings a good sense of immersion and is relatively light-weighted for wearing is illustrated in Fig. 2(b).

Fig. 2. Devices used in system: (a) is the Perception Legacy, (b) is the HTC Vive HMD

3.2 System Architecture

The system architecture is illustrated in Fig. 3. Three computer servers (Mocap servers) are used for data acquisition and data transfer to the stage server. The data from the three Mocap servers are used to update the virtual dancer's motions in a virtual reality environment managed with Unity5 game engine. For a synthesized scene including all the performance elements, we achieve an output frame rate of more than 60 fps.

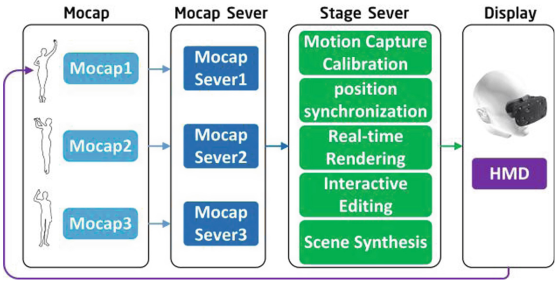


Fig. 3. System architecture.

Three sets of motion capture data were transmitted to the stage sever in real-time, and assigned to three avatars in the virtual stage.

3.3 Multi-person Position Synchronization

As all the motion are captured in different local coordinates, motion calibration and alignments are required to synthesize the virtual dance correctly. Sra and Schmandt presented a full-body tracking system, with two Microsoft Kinect devices to track the full-body of each user and map their movements to an avatar in the virtual world [16]. Kinect was adopted in several research works to capture the dancer's movement [2, 11, 17]. However, its sensor range and FOV are quite limited for this application.

We build the virtual and real space based on the Unity of the "*Left Hand of Coordinate System*". The spatial coordinate system is illustrated in Fig. 4. We can see the $Box(abcd,\ a_1b_1c_1d_1)$ is the full range of activities for dancers in the real space and do not limit the partner and choreographer, the a_1 and c_1 are the position of two optical positioners with regarding the $Plane(abcd)$ as the ground.

We define "K" is the scale factor for the real space and the virtual space:

$$K = \frac{Virtual\_Length}{Real\_Length}$$

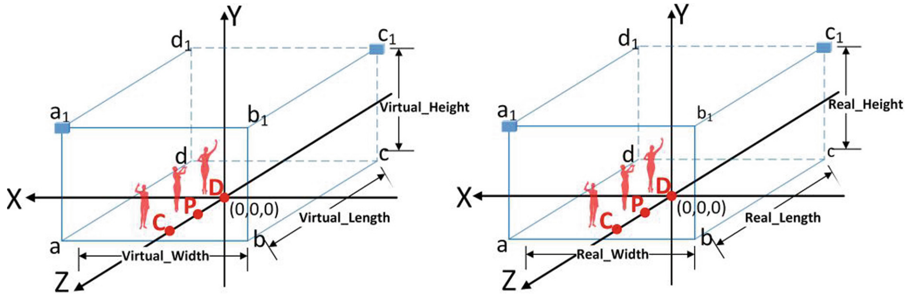


Fig. 4. Position mapping: (a) is the virtual space coordinate system, (b) is the real space coordinate system

We start the position calibration with the initial position of the partner $P_{(position)}$ and the choreographer $C_{(position)}$ according to the positional relationship between the three in the real space with the scale factor "K":

$$P_{(position)}(0, 0, K \times Real\_Dis)$$

$$C_{(position)}(0, 0, 2K \times Real\_Dis)$$

The advantage is that the spatial position of the partner and the choreographer is not constrained by the HTC positioner and is free to move in the performance space thereby increasing the activity space in the dance performance.

3.4 Structure of Virtual Stage

The virtual reality environment is running on a stage sever, providing more functionalities including scene editing and motion capture calibrations. The virtual 3D model of the No. 1 studio of CCTV channel, which is equipped with thousands of square meters of computer controlled LED displays.

The structure of virtual stage is rendered as 3D model is illustrated in Fig. 5. The music, the lighting, the costumes, and the background video shown on the LED displays with certain simplifications to promote the rendering efficiency.

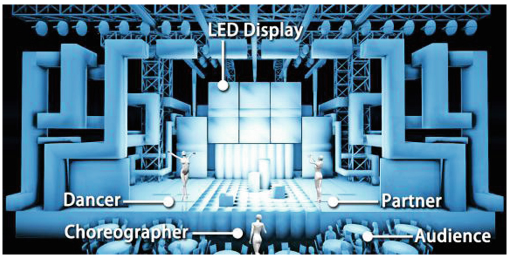


Fig. 5. The structure of virtual stage.

4 Experiment Design

4.1 Experiment Setup

We invited 32 participants to join our experiment and there are 16 females and 16 males. Their age ranges from 20 to 41, and the average age was 26.1 *(SD = 5.38*). Participants played the role of dancer one by one, the partner and choreographer are invited dancing teachers. The movements of the dancers was designed by the dancing teachers and trained before the experiments to master a set of dance movements. All dancers have never used similar virtual reality assistant systems for training and rehearsal. Our experiment space was wide enough to let dancers move in the range of five meters by five meters. We elevated the cable of HMD with a pole to provide the dancer more freedom of motion. Experiment space is illustrated in Fig. 6.

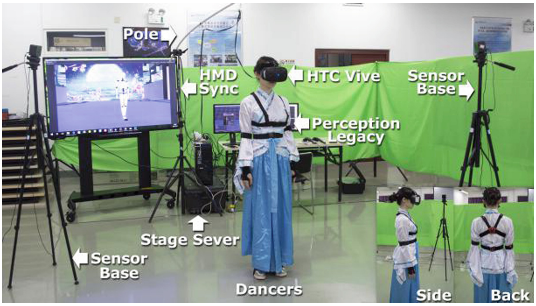


Fig. 6. The experiment space.

4.2 Experiment Process

There are four parts included in our experimental process.

(1) Ready to work: Dancers were informed of the boundary of the experiment, and we had teachers who wear the mo-cap device and HTC Vive to help them.
(2) Position calibration: After the ready, dancers were asked to start the calibration work.
(3) Experiment: Dancers completed this dance in three different view modes. Firstly, in the AV mode, we used a virtual camera to simulate the audience's view, the real-time images captured by the virtual camera transferred to the dancer's HMD, so the dancers could observe their own and partner's performance in the audience's view during the dancing process. The audience's view setup is illustrated in Fig. 7(a). Secondly, in the PV mode, we attached the virtual camera on the partner's head so that the virtual camera moves with the partner and dancers were only able to observe her(his) self-images. The partner's view setup is illustrated in Fig. 7(b). Thirdly, in the CV mode, the choreographer joined in the rehearsal and provided perspective to the dancers with attaching the virtual camera on the choreographer's head. The choreographer's view setup is illustrated in Fig. 7(c).
(4) Interview: Participants would be asked to interviews at the end of the experiment.

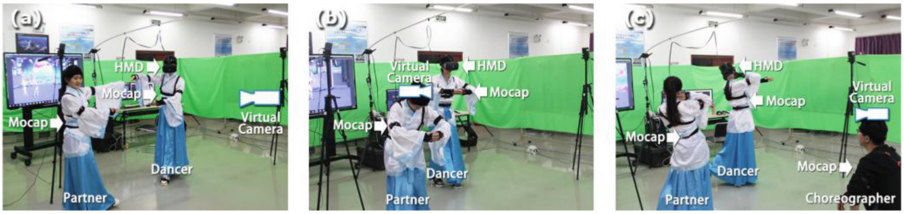


Fig. 7. Three experiment setups in live scenery: (a) audience's view setup, (b) partner's view setup, (c) choreographer's view setup.

5 Experiment

We picked "*The State of Etiquette*" as the dance show for our experiment. The dance reflects the profound Chinese etiquette.

The dance consists of three parts: (1) The primer part: the dancers walks into the center of the stage gently with action stretch, reflecting the Chinese woman's gentle ritual. (2) The climax part: the dancers are inspired by the acceleration of the rhythm. The dancers movements became more enthusiastic with many variations. The dancers and their partners exchanged positions frequently in the stage for more diversive and ornamental formations. (3) The ending part: The dancer's motions are also slowed down with the music, the finally the dancers become still.

In the AV mode, the dancers observed the virtual dance performance, including her (his) own avatar and the partner's avatar through the HMD. The images captured from the audience view transfer to the HMD is illustrated in Fig. 8(a), and the live scenery is in the bottom right corner. In the PV mode, dancers could only see her(his) self-images is illustrated in Fig. 8(b). In the CV mode, it could be observed the dance performances of the dancers themselves and their partner through the perspective of choreographer. Image through the HMD is illustrated in Fig. 8(c).

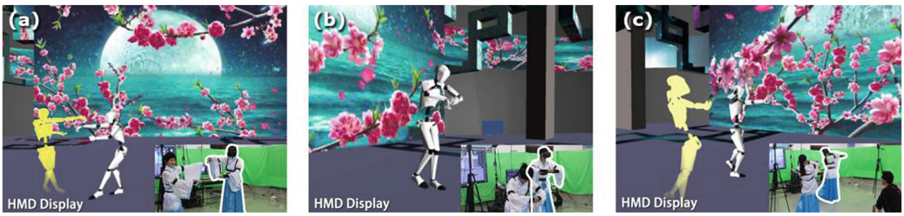


Fig. 8. Rehearsal with three views. Stroke with white in live scenery is the dancer, corresponding with the white avatar in HMD display: (a) audience's view, (b) partner's view, (c) choreographer's view

6 Result and Discussion

After the interview, we got a lot of useful feedback, and we described from three aspects respectively.

(1) Immersion: Most of the participants said they had a better sense of immersion. The synthesis effect of the virtual stage was helpful for dance rehearsal and the position of the dancer on the virtual stage was more accurate.
(2) View adaptability: Participants said they did get an unprecedented experience by the three perspectives. In addition, the observation of the dance performance is more intuitive in the AV mode. The PV and CV mode was relatively poor compared to the AV mode. In the PV mode, when the distance between the dancers very close, the vision angle of the participants occasionally penetrated their own avatars in the virtual stage. With the CV, the participants could have a certain sense of the creative intentions of the choreographer.
(3) System practicality: This system can provide a valuable reference for dance rehearsal and it can be used in other forms of performance such as gymnastics show, opera and stage show, etc.

7 Conclusion and Future Work

The purpose of this work is help the dancers integrate their own dance movement with other performance elements. A preliminary test shows that the proposed system provided third-person views for dancers and improve their understanding of: (1) augment dancer's perception of their dance performance; (2) their understanding of the choreography; and (3) evaluate the dancer's performance standing on a different identity. The proposed system has potential to become a powerful rehearsal tool for many forms of live performances. It would be interesting for the audience to join the live performance and watch the show from the perspective of the dancers.

Acknowledgments. The research leading to these results has received funding from National Natural Science Foundation of China under Grants No. 60975013, National Natural Science Foundation of China No. 61202243 Research on Realtime Motion Capture with Minimal Inertial Sensors and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028.

References

1. Azarbayejani, A., Wren, C., Pentl, A.: Real-time 3-D tracking of the human body. In: Proceedings of IMAGE'COM 1996, pp. 780–785 (1996)
2. Alexiadis, D.S., Kelly, P., Daras, P., et al.: Evaluating a dancer's performance using kinect-based skeleton tracking, pp. 659–662 (2011)

3. Anderson, F., Grossman, T., Matejka, J., Fitzmaurice, G.: YouMove: enhancing movement training with an augmented reality mirror. In: Proceedings of UIST 2013, pp. 311–320. ACM, New York (2013)
4. Chan, J.C.P., Leung, H., Tang, J.K.T., et al.: A virtual reality dance training system using motion capture technology. IEEE Trans. Learn. Technol. **4**(2), 187–195 (2011)
5. Chua, P.T., Crivella, R., Bo, D., et al.: Training for physical tasks in virtual environments: Tai Chi. In: IEEE Virtual Reality, p. 87. IEEE Computer Society (2003)
6. Ho, E., Chan, J., Komura, T., Leung, H.: Interactive partner control in close interactions for realtime applications. ACM Trans. Multimed. Comput. Commun. Appl. **9**(3), 21 (2013)
7. Hallam, J., Keen, E., Lee, C., et al.: Ballet hero: building a garment for memetic embodiment in dance learning. In: Proceedings of the 2014 ACM International Symposium on Wearable Computers: Adjunct Program. pp. 49–54. ACM (2014)
8. Hachimura, K., Kato, H., Tamura, H.: A prototype dance training support system with motion capture and mixed reality technologies. In: IEEE International Workshop on Robot and Human Interactive Communication, pp. 217–222. IEEE (2004)
9. Han, P.H., Chen, K.W., Hsieh, C.H., et al.: AR-Arm: augmented visualization for guiding arm movement in the first-person perspective. In: Augmented Human International Conference, p. 31. ACM (2016)
10. Han, P.H., Chen, Y.S., Zhong, Y., et al.: My Tai-Chi coaches: an augmented-learning tool for practicing Tai-Chi Chuan. In: Augmented Human International Conference, p. 25. ACM (2017)
11. Kitsikidis, A., Dimitropoulos, K., Douka, S., et al.: Dance analysis using multiple kinect sensors. In: International Conference on Computer Vision Theory and Applications, pp. 789–795. IEEE, Hfh (2014)
12. Kyan, M., Sun, G., Li, H., et al.: An approach to ballet dance training through MS kinect and visualization in a CAVE virtual reality environment. ACM Trans. Intell. Syst. Technol. **6**(2), 291–300 (2015)
13. Uejou, M., Huang, H.-H., Lee, J.-H., Kawagoe, K.: Toward a conversational virtual instructor of ballroom dance. In: Intelligent Virtual Agents, pp. 477–478. Springer, Berlin (2011)
14. Marquardt, Z., Beira, J., Em, N., Paiva, I., Kox, S.: Super Mirror: a kinect interface for ballet dancers. In: Proceedings of CHI 2012 Extended Abstracts, pp. 1619–1624. ACM, New York (2012)
15. Niwayama, T., Nakamura, A., Tabata, S., et al.: Mobile robot system for easy dance training. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, vol. 3, pp. 2223–2228. IEEE Xplore (2004)
16. Sra, M., Schmandt, C.: MetaSpace: full-body tracking for immersive multiperson virtual reality. In: Adjunct Proceedings of the ACM Symposium on User Interface Software & Technology, pp. 47–48. ACM (2015)
17. Semwal, S.K., Hightower, R., Stansfield, S.: Mapping algorithms for real-time control of an avatar using eight sensors. Presence: Teleoper. Virtual Environ. **7**(1), 1–21 (1998)
18. Trajkova, M., Cafaro, F.: E-ballet: designing for remote ballet learning. In: ACM International Joint Conference, pp. 213–216 (2016)
19. Yan, S., et al.: OutsideMe. In: The ACM Conference Extended, pp. 965–970. ACM (2015)
20. Yang, U., Kim, G.J.: Implementation and evaluation of "just follow me": an immersive, VR-based, motion-training system. Presence: Teleoper. Virtual Environ. **11**(3), 304–323 (2002)
21. Myroniv, B., Wu, C.-W., Ren, Y., Christian, A.B., Bajo, E., Tseng, Y.-C.: Analyzing user emotions via physiology signals. Data Sci. Pattern Recogn. **1**(2), 11–25 (2017)

Augmented Reality-Based Pre-visualization for Film Making

Tong Xue[1], Gangyi Ding[1], and Fuquan Zhang[1,2](✉)

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
{xuetong, dgy}@bit.edu.cn, 8528750@qq.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350108, China

Abstract. With the development of the augmented reality technologies, wearing devices such as Microsoft Hololens and Google glass become highly popular recently. In this paper, we proposed a novel augmented reality application that helps the film directors to complete and facilitate their film rehearsal process. We simulated a design space to provide film directors with a high-fidelity visual-audio experience. The directors wearing the hololens are allowed to observe the synthesized virtual-reality space in which they can recognize the whole design space, prefabricated boxes, and modify the space by defined gestures. Feedbacks from directors indicate the proposed tool is effective and the preliminary experimental results agree with our observations. In general, our research offers the guidelines for novel film rehearsal tactics.

Keywords: Augmented reality · Pre-visualization · Film making
Subjective evaluation

1 Introduction

Before a film starts shooting, pre-visualization is essential [2]. On one hand, it helps the film directors identify the relationship among the performance elements in the scene. On the other hand, directors could intuitively understand the interaction between the actors and the performance elements. What's more, filmmakers may pay more attention to the film itself, since pre-visualization saves much time and reduce costs (Fig. 1).

We present a two-dimensional pre-visualization model, which could be used to describe research on this field. In general, film making is done by the field shooting. Research to date has tended to explore the area of using digital technology to reconstruct and pre-visualize performing arts events [6], such as scenic design, lighting and sound effects, actors. OutsideMe system [4] augmented external self-image by display the dancer's self-image captured by a wide-angle camera using Oculus Rift HMD. The same method was used to design a training system for ballet [3]. OutilNum project [1] allowed a real actor to pre-visualize virtual scenes and to act with virtual actors, mainly using Oculus rift HMD, Microsoft Kinect and Natural Motion OptiTrack. The above-mentioned rehearsal systems all focus on actors' pre-visualization with the help of virtual reality devices. In these cases, users cannot move freely in space with a wired

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 217–222, 2019.
https://doi.org/10.1007/978-3-030-04585-2_26

equipment. Also, the interaction between the users and the scenes is very limited. Recently, a small number of researchers try to use augmented reality plugins to mix virtual and reality. For instance, Stamm et al. [5] presents a low-cost system that uses Tango to pre-visualize virtual assets in on-set film production in order to help directors perceive environment of the production. Since Project Tango is Android-based, which has some disadvantages like short battery life, limited use distance, weak immersion and just for directors, we need to develop a new application for pre-visualization of film making, to better integrate virtual and reality, while considering both the directors and the actors.

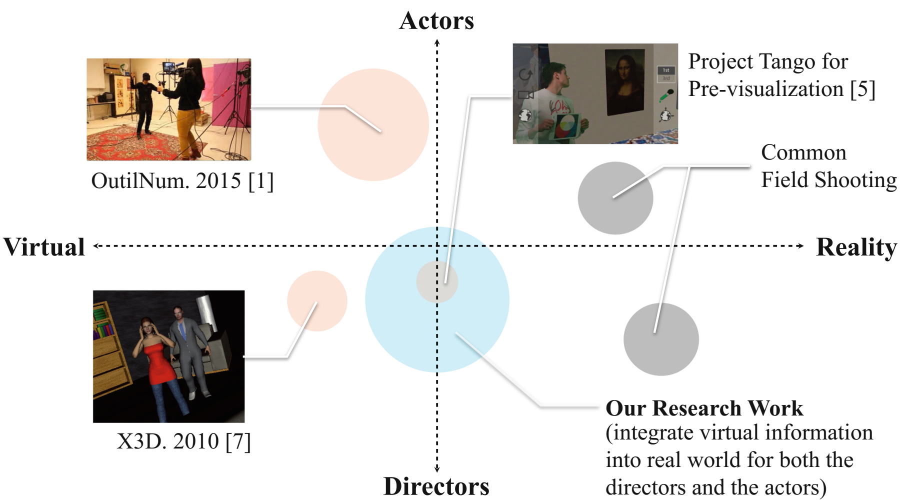


Fig. 1. A two-dimensional pre-visualization model. Some typical research and our work are listed in the plot.

Against this broad backdrop, we use the hololens to bridge the gap, accomplishing a new film rehearsal system. Microsoft HoloLens is such a device that integrates virtual information such as objects, videos, pictures into the real world, so that building a brand new scene [10]. It enables users to engage with their digital contents and interact with the holograms in the world around them. In our application, directors can select objects from prefabricated prop library, actor library, action library and place them within the physical world. This proposed system can be used inside the studio or an ordinary room, which provides an unprecedented perspective to the film directors and help them make a film more efficiently and thoroughly.

2 Methodology

2.1 Implementation

Our system consists of four layers (Fig. 2):

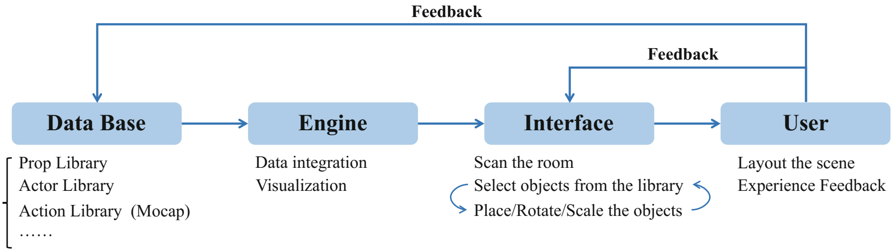


Fig. 2. Overview of our system.

(1) The Data Base layer: Before the system development, we first prepared the data base. It contains the prop library, the character library, the action library and so on. We have collected 50 kinds of common props in film making, each with 10–20 different styles. The action clips in the action library were captured from professional actors.

(2) The Engine layer: The goal of this layer is to drive data and implement logic. We integrated the whole elements, visualized these models of the objects and then deployed our application in the hololens. The system was developed using the Unity framework with C# scripts.

(3) The Interface layer: This layer integrates spatial scanning for high quality, high precision reconstruction of the real room. We also defined gaze and specific gestures for interacting with virtual objects. For example, a one-finger tap acts like clicking button allows things selected, while sliding fingers could make the selected objects rotate.

(4) The User layer. Wearing the hololens, users could see virtual objects and layout the scene according to their needs. At the end of the experience, the researchers had the participants remove their headset and fill out a post-experiment questionnaire to receive subjective feedback of our system, which would help us modify the database and the interactive modes.

2.2 Experiment

The purpose of this research is to help filmmakers pre-visualize their film-video works. Thus, to determine director's acceptance and usability of our system, we conducted a pilot user study (Fig. 3).

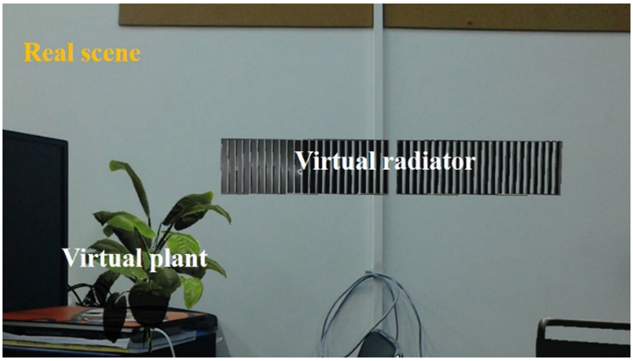


Fig. 3. A screenshot of a scene in the layout. The participant placed multiple holograms including plant and radiator to the scene of his work.

The 45 participants consist of students and teachers, whose major is film and television production, and independent filmmakers ranging from 20 to 40 years of age, to represent various skill level of people in the field of film making. Among these participants, the number of students, teachers and filmmakers is 15. All the participants were asked to construct an original film scene wearing the hololens. After the experience, each participant completed a questionnaire with 20 questions. Also, they were asked if they would like to voluntarily participate in a focus group discussion.

3 Results and Discussion

In post-experiment questionnaire, we applied 5-point Likert Scale (5 = totally agree, 1 = totally disagree). The 20 questions are divided into four subjects: Functionality, Immersion, Learnability and Efficiency [8]. As shown in Table 1, we clearly observed that there were no factors dipping below the neutral threshold, indicating positive feedback of our proposed system. From the functional analysis of the results, the difference between students and filmmakers is obvious. We attribute this discrepancy to the limitations of prefabricated objects, while filmmakers may focus more on the ultimate effect. In the aspect of immersion, the mean scores of participants in different categories were quite close but not as high as other indicators. Viewing the scene through special lenses would distract the users' attention to some extent. About learnability and efficiency, the participants basically agreed that they can easily interact with the application based on friendly gestures. During the focus group, the participants were asked questions related to their experience and proposals. Several participants reported dizziness, especially after wearing a long time.

Table 1. Mean and standard deviation for overall diagnostic factors scores.

Factors	Students		Teachers		Film makers	
	Score	SD	Score	SD	Score	SD
Functionality	4.81	0.25	4.73	0.25	4.15	0.34
Immersion	4.52	0.39	4.53	0.56	4.53	0.23
Learnability	4.78	0.21	4.76	0.32	4.66	0.55
Efficiency	4.87	0.33	4.89	0.19	4.95	0.37

Consistent with other virtual reality pre-visualization systems, the original intention is to simulate the movie scenes and storylines in the early stages of film making [2]. However, instead of virtual elements in virtual environment, our rehearsal tool creatively allows directors adding virtual performing elements to the realistic acting space flexibly. What's more, we provide filmmakers with substantial prefabricated objects placed in different libraries by type, including prop library, actor library and action library. The most striking aspect of our research is that directors can walk around in the real room and interact with visible virtual and real objects without any restrictions. Therefore, users would have a stronger sense of immersion, which is more conducive to film making. Also, actors could experience the rehearsal from the perspective of directors and have a better understanding of the production in advance.

Since the device is still in development and just a prototype version, a few limitations like a relatively narrow field of view, overheating and a little bit heavy may make users feel discomfort during extended periods of use. Moreover, the first step is to scan the users' environment based on walls, ground and ceilings, so it's hard to use this system outdoors. As for the evaluation phase, it revealed strength and weakness of the system. However, both of the questionnaire and the focus group are subjective evaluation tools, which are difficult to get purely honesty feedback. And we didn't record the actors' feedback.

4 Conclusion and Future Work

In this work, the prototype of AR system for pre-visualization is presented. A preliminary test shows that the proposed application has a great usability and can significantly improve film making efficiency for both student directors and independent directors.

In our future work, we try to improve the usability of the application by enriching the types of prefabricated objects, increasing the framerate of scene rendering, and reducing the weight of the wearable devices. And the application can be extended to directly generate 3D virtual performance content in the real space and become a novice form of performing art, by using the deep learning algorithm [9]. It would also be interesting for the actors and audience to join the live performance and experience the show from the perspective of directors. In addition, we will address our evaluation shortcomings as the project continues by exploring biometric evaluation techniques (Fig. 4).

Fig. 4. A participant interacting with virtual object in the library.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204).

References

1. Plessiet, C., Chaabane, S., Khemiri, G.: Autonomous and interactive virtual actor, cooperative virtual environment for immersive Previsualisation tool oriented to movies. In: Virtual Reality International Conference, pp. 1–4 (2015)
2. Nitsche, M.: Experiments in the use of game technology for pre-visualization. In: Conference on Future Play: Research, Play, Share, Future Play, vol. 3, pp. 160–165 (2008)
3. Trajkova, M., Cafaro, F.: E-ballet: designing for remote ballet learning. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, pp. 213–216 (2016)
4. Yan, S., Ding, G., Guan, Z., Sun, N., Li, H., Zhang, L.: OutsideMe: augmenting dancer's external self-image by using a mixed reality system. In: ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 965–970 (2015)
5. Stamm, A., Teall, P., Benedicto, G.B.: Augmented virtuality in real time for pre-visualization in film. In: IEEE 3D User Interfaces, pp. 183–186 (2016)
6. Yan, S., Ding, G.Y., Li, H., Sun, N., Wu, Y., Guan, Z., et al.: Enhancing audience engagement in performing arts through an adaptive virtual environment with a brain-computer interface. In: ACM Proceedings of the 21st International Conference on Intelligent User Interfaces, pp. 306–316 (2016)
7. Jung, Y., Wagner, S., Jung, C., Behr, J., Fellner, D.: Storyboarding and pre-visualization with X3D. In: ACM International Conference on Web 3D Technology, pp. 73–82 (2010)
8. Kalawsky, R.S.: VRUSE—A computerised diagnostic tool: for usability evaluation of virtual/synthetic environment systems. Appl. Ergon. **30**(1), 11–25 (1999)
9. Chen, X.Y., Peng, X.Y., Li, J.B., Peng, Y.: Overview of deep kernel learning based techniques and applications. J. Netw. Intell. **1**(3), 83–98 (2016)
10. Microsoft HoloLens Homepage. https://www.microsoft.com/zh-cn/hololens

MR-APG: An Improved Model for Swarm Intelligence Movement Coordination

Hui Yang[1], Linlin Ci[3], Fuquan Zhang[1,2(✉)], Minghua Yang[3], Yu Mao[1], and Ke Niu[4]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
zlj-1943@163.com, maoyu_bit@163.com, 8528750@qq.com
[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, China
[3] RocketMilitary Equipment Research Institute, Beijing, China
cilinlin@263.net, yang-mh@foxmail.com
[4] Computer School, Beijing Information Science and Technology University, Beijing, China
niuke@bistu.edu.cn

Abstract. Swarm intelligence based on the bionics have become an emerging research field, and researchers have proposed some kinds of intelligent models, such as ant colony, free group wolves, fish flock, insects group, bacteria and bacteria immune etc. The behavior of common features is: the individual behavior and function is simple, but a rally for the swarm after a complex cluster behavior ability. In mobile swarm group individuals within a certain space can keep a certain distance to avoid collision in order to group coordinated movement. In this paper, an attempt has been made to optimize the continuous and collaborative of the swarm, when the swarm nodes in the process of moving and turning. This paper puts forward the Multi-Restrained Artificial Pigeon Group (MR-APG) model based on multiple constraint factors, and introduce the concepts of quantitative coordination of nodes to optimize the nodes "over-dispersion" or "over-assembly" phenomenon. The experimental results show that, the higer α_1 value will lead to average movement synchronization ratio μ_{sync} of MR-APG model will be lower. The results also show that, comparing with the traditional A/R model ($\mu_{sync} = 73\%$) and 3D-VFA model ($\mu_{sync} = 84\%$), the proposed MR-APG method can make μ_{sync} equal to 95%, moreover, its only takes half time of A/R model to make μ_{sync} equal to 80%.

Keywords: Swarm intelligence · Group synchronization
Quantitative movement

1 Introduction

Swarm intelligence, as a scientific discipline including research fields such as swarm optimization or distributed control in collective individuals, was born from biological insights about the incredible abilities of social insects to solve their everyday-life problems. Its model based on the bionics have become an emerging field of research in

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 223–230, 2019.
https://doi.org/10.1007/978-3-030-04585-2_27

recent times, such as ant colony, the study found that a lot of free group wolves, fish, insects, group, such as bacteria and bacteria immune clustering behavior of common features are: individual behavior and the function is simple, but a rally for the group after a complex cluster behavior ability. In a mobile group of individuals within a certain space can keep a certain distance to avoid collision; at the same time, when an individual node limited damage or attacks by predators, group still can maintain its dynamic characteristics.

In this paper, we optimize the continuous and collaborative of the pigeon swarm, when the nodes in the process of moving and turning. Inspiring by the A/R standard model, this paper puts forward the model of artificial deterrent based on multiple constraint factors, such as Near Neighbor Force $\overrightarrow{f_{ij}}$, Uniform Force UF_{ij}.and proposed pigeon movement direction factor ϕ and pigeon movement speed factor Π to quantize the state of the moving pigeon node. Then we build a more optimized model than A/R model and 3D-VFA model. Multi-Restrained Artificial Pigeon Group (MR-APG) model focused on the coordination problem of transport fleet, introduced the concept of quantitative coordination, to optimize the nodes "over-dispersion" or "over-assembly" phenomenon, and can ensure the nodes in the swarm group maintain dynamic consistency, and avoid node colliding. MR-APG model only takes half time of A/R model to achieve 80% average synchronization ratio, which can meet the requirement of high real-time swarm coordination in some particular scenarios.

2 Related Works

Even though we may not fully understand the true mechanisms that lead to the self-organization and intelligent characteristics of a complex system, researchers have successfully developed optimization algorithms based on swarm intelligence.

At the beginning, Reynolds [2, 11] put forward cluster motion simulation program for the first time, and successfully simulated the birds flying, aggregation and walls in 1987. Each bird by an agent, the agent through its own neighborhood within the scope of environmental information. Ros and Bassman [7] used the characteristics of biological cluster early computer simulation is carried out cluster behavior of life, and put forward the famous Boid model, the separation of cluster motion, calibration and aggregation of three basic rules, became the milestone group deployment of follow-up studies.

2.1 Virtual Force Field

The basic idea of virtual force field is constructing the target orientation gravitational field and repulsive force field around the virtual force field artificially, every nodes along the virtual repulsive force and gravity force direction of movement.

Gazi and Passino [3] separation based on Boid Model (BM) and close to the rules (A/R) is designed to attract/rejection model, and using this model, they proved that the random node distribution of cases, individuals within the limited time to form a ring, and the upper bound of the radius of the ring is given. Then Park and Han [8] proposed

3D-VFA model in order to resolve some application in 3-dimensions, which introduced 3-dimensions descriptor and node accelerated speed.

Partridge et al. [4] improved Vicsek model of linear motion direction angle of the update function, points out that when individuals generally consistent direction coordinated movement, will reach the population migration synergy group of nodes. Couzin of Princeton university and others more universal group synergy model is put forward, it absorbs the Attract/Reject (A/R) model and Vicsek model, the advantages of the proposed Three Circle Model (TCM).

2.2 Neighbor Information Reference

Attanasi et al. [5] published hundreds of birds against doing large sample test data in Nature (2014), the paper proposed that the hierarchy of birds in the main function of the data rapidly flooding, although the experiment using larger sample sizes, but lacking of detailed analysis for eachbird. Zhang [12] figure out a mixture strategies which birds flying in collaborative process: “reference leader” and “refrencen eighbours”. When the flight path “smooth” (turning curvature is less than a certain threshold), the birds with reference to the surrounding neighbors, the average flight direction.

Dorigo and Maniezzo [10] puts forward a kind of group formation control optimization strategy by adjusting speed by avoiding nodes gathered near the collision, through simulation experiment to investigate how collaboration under limited resources to achieve guide most nodes synergy effect with a small number of nodes method.

2.3 ACO and PSO

The ACO (Ant Colony Optimization) meta-heuristics is inspired by the foraging behaviour of ants colony, and the PSO (Particle Swarm Optimization) meta-heuristics is motivated by the swarm coordinate movement of birds and fish group. Both methods above introduced have been applied successfully in a vast range of problems [3]. In recent times, several swarm intelligence algorithms have been presented, inspired by Bacterial Foraging [11], fireflies bio-luminslime life cycle [13].

3 MR-APG Model

3.1 Force Field Concepts

As the standard of the A/R model is relatively simple, and does not consider the node perception scope, communication parameters etc. Based on the basic idea of the A/R model, This section propose Multi-Restrained Artificial Pigeon Group (MR-APG) model, and the related concepts are defined as follows:

Definition 1. The Near Neighbor Force $\overrightarrow{f_{ij}}$: $\overrightarrow{f_{ij}}$ between node i and its neighbor node j can be calculated by the formula as fellow:

$$\left|\overrightarrow{f_{ij}}\right| = ((d(i,j) - R_b) * \delta)^k \tag{1}$$

In formula 3, $d(i,j)$ represent Euclidean space distance, k is gain factor, which enlarge the influence of the node near radius edge. Adjusting the value of k, we can control the swarm pigeons density degree. Generally, $\delta \in (5, 10)$ is used to guarantee the force the derivative of greater than zero. Therefore, according to the node by force, make the neighbor nodes gathered or separated.

Definition 2. Uniform Force UF_{ij}: for any pigeon node i and its neighbors, we can calculate UF_{ij} neighbor nodes of the size of the uniform force calculated by the formula as fellow:

$$\left|\overrightarrow{UF_{ij}}\right| = ((d(s_i, s_j) - \overline{D_i}) * \beta)^k \tag{2}$$

Using formulas 1 and 2, we can figure out the net force $\overrightarrow{F_i}$ which the pigeon node i suffered, and it can be represented by formula 4.

$$\overrightarrow{F_i} = \alpha_1 f_{A/R}(x) + \alpha_2\left|\overrightarrow{f_{ij}}\right| + \alpha_3\left|\overrightarrow{UF_{ij}}\right| \tag{3}$$

In formula 3, α_i is weight factor, generally, $\alpha_1 > 0.5$, and $\sum_{i=1}^{3}\alpha_i = 1$.

3.2 Quantification of Direction Synchronization

Because standard A/R model does not conform to the real node objects of resultant force, which could directly lead to the change in the velocity of swarm nodes. In order to quantize any one pigeon node in swarm, this section defines pigeon movement direction factor ϕ. The node's direction synchronization formula is as follows:

$$\begin{cases} E[\phi_i(t)] = \sum_{j \in N_i} \left(1 - \theta_i(t) \cdot \theta_j(t)\right)/N_i \\ \phi_{\min i}(t) = \min_{j \in N_i}\left(1 - \theta_i(t) \cdot \theta_j(t)\right) \end{cases} \tag{4}$$

In this case, t is time variable, N_i is the set of neighbors of i where $|N_i|$ is the number of neighbors, the movement angle $\theta_i(t) = [\cos\theta, \sin\theta]^{\mathrm{T}}$, and the symbol '.' denotes 'inner product'. Hence, $1 - \theta_i(t) \cdot \theta_j(t)$ represents the "error" or "difference" between the direction of pigeon individual i and its neighbor j. Specifically, $1 - \theta_i(t) \cdot \theta_j(t) = 0$ presents a couple with fully coordinated movements and $1 - \theta_i(t) \cdot \theta_j(t) = 2$ presents a couple with opposite movements. Respectively, $\phi_{\min i}$ denotes the average and minimal values. Respectively, the node movement differences of the directions between pigeon individual i and its neighbor j.

3.3 Quantification of Speed Synchronization

In the same manner, this section will introduce quantification speed information, and defines pigeon movement speed factor Π. The node's speed synchronization formula is as follows:

$$\begin{cases} E[\Pi_i(t)] = \left(\sum_{j \in N_i} \frac{\left| \|v_i(t)\| - \|v_j(t)\| \right|}{\|v_i(t)\|} \right) \frac{1}{N_i} \\ \Pi_{\min i}(t) = \min_{j \in N_i} \frac{\left\| v_i(t) - v_j(t) \right\|}{\|v_i(t)\|} \end{cases} \tag{5}$$

In this case, t is time variable, N_i is the set of neighbors of node i. $E[\Pi_i(t)]$ represents the expectation of speed difference between node i and its surrounding neighbors. Correspondingly, $\Pi_{\min i}(t)$ represents minimal speed difference between node i and its surrounding neighbors j. $\|\cdot\|$ is norm operator, which denotes the speed value of the node.

3.4 Average Movement Synchronization Ratio

According to Vicsek model [1], we introduce and improve the average movement synchronization ratio μ_{sync} as fellows:

$$\mu_{sync} = \sum_{t=1}^{T} \frac{\left\| \sum_{i \in N} v_i(t) \right\|}{n \mathrm{T} v_0(t)} \tag{6}$$

Obviously, μ_{sync} is due to the size of the node speed changes with factor t, when the $\mu_{sync} = 1$, means node i is fully synchronous speed with its neighbors overall, and when $\mu_{sync} = 0$, means node i the speed change completely random compared to its neighbors.

3.5 Experiment Settings

In this paper, we using Matlab2011b simulates the swarm pigeon group movement, At the beginning, the experiment setup for 2 scenarios: 2-dimensional scenario and 3-dimensional scenario, then the detailed information is shown in Table 1.

Table 1. Experiment settings

Scenario	Number of nodes	Time steps	Initialization mode	k	σ	β
2-D simulation	30	200	Random	0.4	6	2
3-D simulation	40	200	Random	0.35	7	2

For the sake of simplicity, we omit the node's perception of the radius of the ball in the figure, and the red dot represents the location of body section idea, blue "tail" represents the movement direction of the node.

3.6 Simulation

Figure 1 shows the initialized nodes state in 2-dimensional scenario, at the beginning, the direction and the speed of 30 nodes are random initialized, and each node begin to move by the net force $\overrightarrow{F_i}$ which defined in MR-APG model, when t = 200, the nodes movement direction is almost consistent.

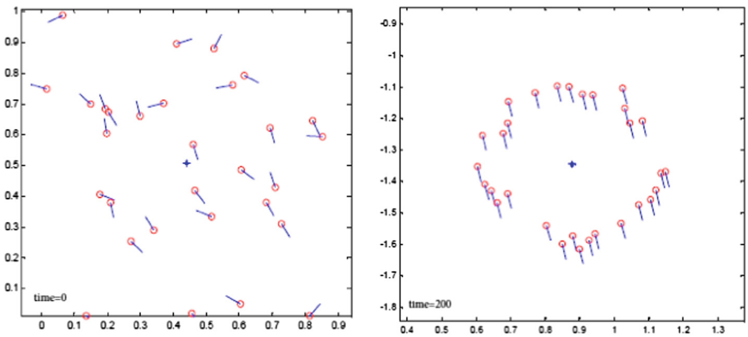


Fig. 1. Quantification of movement synchronization (2-D)

Figure 2 shows the initialized nodes state in 3-dimensional scenario, at the beginning, the direction and the speed of 40 nodes are random initialized, and each node begin to move by the net force $\overrightarrow{F_i}$ which defined in MR-APG model, when t = 200, the nodes movement direction is almost consistent in 3-dimensional scenario.

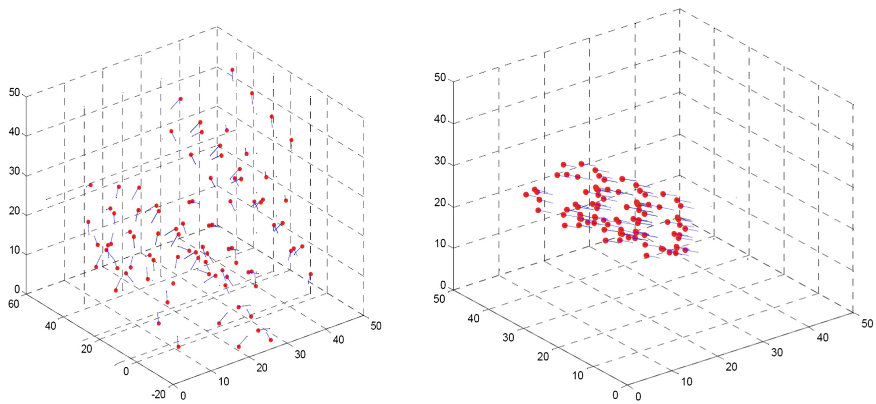


Fig. 2. Quantification of movement synchronization (3-D)

3.7 Experiment Analysis

In Sect. 3.6, we have simulated 30 nodes movement synchronization in two scenarios respectively in Matlab.

Figure 3(a) shows the average movement synchronization ratio of MR-APG with different weigh factor α_i. When t = 40 per unit time, the synchronization rate μ_{sync} climbed stablely, which means the node *i* can adjust the force. Apparently, in the condition of $\alpha_1 > 0.5, \alpha_2 = \alpha_3$, the higer α_1 value will lead to promotion of A/R model importance, which means MR-APG model will be more similar to A/R model, without some necessary restraint factors, synchronization rate the lower.

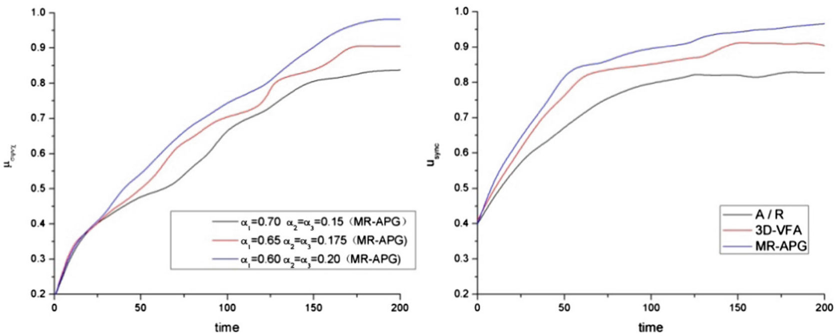


Fig. 3. (a) MR-APG with different weight factors (b) Algorithms comparison

As shown in Fig. 3(b), the change of average movement synchronization μ_{sync} during the experiment process as shown in Fig. 3(b). At the beginning, the speed of each node is initialized random size (a reasonable range). When t = 30 per unit time, the speed synchronization rate climbed almost exponentially with A/R, 3D-VFA and MR-APG algorithm, which means the node *i* can adjust the force quickly. And then the A/R model increase to 75% (its limit) very slowly, when t = 50 per unit time, the speed of node to reach the basic synchronization, and always keep pace. Comparing with 3D-VFA, MR-APG can reach higher synchronous ratio, when all nodes speed same size, direction and size from below the speed synchronization degree change, when t = 100 per unit time, the movement synchronization rate even stabilizes above 90%. In addition, we also noticed that μ_{sync} couldn't continue to arise in figure after t = 150, which is due to the time delay of the model among the swarm nodes.

4 Conclusion

This paper proposed MR-APG model in order to optimize the continuous and collaborative of the swarm in the process of moving and turning. Our proposed model includes multiple constraint factors, and firstly introduces the concepts of quantitative coordination of nodes to optimize the nodes "over-dispersion" or "over-assembly" phenomenon.

The experimental results show that, the higer α_1 value will lead to average movement synchronization ratio μ_{sync} will be lower, and the experimental results also show that, compared with the traditional A/R model and 3D-VFA model, the proposed MR-APG method can make $\mu_{sync} = 95\%$, and MR-APG model only takes half time of A/R model to achieve 80% average synchronization ratio, which can meet the requirement of high real-time swarm coordination in some particular scenarios.

This research is supported by the National natural science foundation of China (grant no. 61063042), the State Key Laboratory of Digital Publishing Technology.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028.

References

1. Vicsek, T., Zafeiris, A.: Collective motion. Phys. Rep. **517**, 71–140 (2012)
2. Lukeman, R., Li, Y.X., Edelstein-Keshet, L.: Inferring individual rules from collective behavior. Proc. Natl. Acad. Sci. U.S.A. **107**, 12576–12580 (2010)
3. Gazi, V., Passino, K.M.: Stability analysis of swarms. In: Proceedings of American Control Conference, Anchorage, Alaska, pp. 1813–1818, May 2002
4. Partridge, B.L., Pitcher, T., Cullen, J.M., Wilson, J.: The three-dimensional structure of fish schools. Behav. Ecol. Sociobiol. **6**(4), 277–288 (1980)
5. Attanasi, A., et al.: Information transfer and behavioural inertia in starling flocks. Nat. Phys. **464**(1290), 890–899 (2014)
6. Breder, C.M.: Equations descriptive of fish schools and other animal aggregations. Ecology **35**(3), 361–370 (1994)
7. Ros, I.G., Bassman, L.C., Badger, M.A.: Pigeons steer like helicopters and generate down- and upstroke lift during low speed turns. Proc. Nat. Acad. Sci. **45**(7) (2011)
8. Park, H., Han, J.-H., Kim, J.-H.: Swarm intelligence-based sensor network deployment strategy. In: 2011 IEEE Congress on Evolutionary Computation (2010)
9. Locatelli, M., Raber, U.: Packing equal circles in a square: a deterministicglobal optimization approach. Discrete Appl. Math. **122**(1), 139–166 (2012)
10. Dorigo, M., Maniezzo, V., Colorni, A.: Ant system: optimization by a colony of cooperating agents. IEEE Trans. Syst. Man Cybern. Part B **26**(1), 29–41 (2009)
11. Reynolds, C.W.: Flocks, herds and schools: a distributed behavioral model. ACM SIGGRAPH Comput. Graph. **21**(4), 25–34 (1987)
12. Zhang, H.T., Chen, M.Z.Q., Zhou, T., et al.: Synchronization and Aggregation of Self-propelled Particles via Partial Communication arXiv:0707.3402 (2013)
13. Tian, B.-M., Yang, H.-X., Li, W., et al.: Optimal view angle in collective dynamics of self-propelled agents. Phys. Rev. E **79**(5), 052102 (2009)
14. Peng, L., Zhao, Y., Tian, B., et al.: Consensus of self-driven particles with avoidance of collisions. Phys. Rev. E **79**, 026113 (2009)
15. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., Shochet, O.: Novel type of phase transition in a system of self-driven particles. Phys. Rev. Lett. **75**(6), 1226 (1995)
16. Nagy, M., Akos, Z., Biro, D., Vicsek, T.: Hierachical group dynamics in pigeon flocks. Nature **464**, 890–893 (2010)

Regional Coverage Monitoring Planning Technology for Multi-UAV Based on Pruning PSO

KaiXuan Wang[1], YuTing Shen[2], FuQuan Zhang[3,4](✉), Zhuo Liang[1], ZhiGuo Song[1], and YanPeng Pan[1]

[1] China Academy of Launch Vehicle Technology, Beijing 100076, People's Republic of China
kingkaixuan@126.com, liangzhuo_nust@163.com, songzh166323@163.com, panypxgdxgd_4@163.com

[2] The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050000, People's Republic of China
rebecca1224kaoyan@163.com

[3] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, People's Republic of China
8528750@qq.com

[4] Digital Performance and Simulation Technology Lab, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China

Abstract. In recent years, with the application and gradual popularization of UAV technology in many fields, the normalization of UAV aerial photography has become a common phenomenon. There are few studies on Multi-UAV regional coverage monitoring. In this study, PSO algorithm based on pruning can be used to solve optimization problem, the multi-agent technology solve the problem of UAV group for self-perception and decision-making, the UAV group, namely multi-agent particles, can automatically initialize the perception agent particle whether meet the requirements of the iteration, whether to need to be pruned, and decide to optimize the iteration or increase the agent number, Finally, simulation verification shows that the optimized algorithm can finish the work faster under the premise of guaranteeing the effect.

Keywords: UAV · Pruning · PSO · Regional coverage monitoring

1 Preface

In recent years, with the application and gradual popularization of UAV technology in many fields, the normalization of UAV aerial photography has become a common phenomenon. The UAV is low-cost, easy to carry, simple to operate, and has a wide range of applications. UAV has great potential and development prospects, especially for its ability to enter a place where people cannot enter in natural disasters or emergencies quickly and safely, In this paper, the multi-UAV aerial photography area planning technology based on Pruning PSO algorithm and Agents is used to realize the rapid planning and location optimization of multi-UAV with fast and reasonable multi-UAV location distribution, to achieve continuous dynamic observation of the required

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 231–238, 2019.
https://doi.org/10.1007/978-3-030-04585-2_28

detection area, and support various fields such as news production, geological disaster observation [1], traffic observation, and UAV self-organizing network [2].

In the research of the distributed motion model of the UAV group, it is a common consensus of researchers in various countries that multi-UAV can work better by collaborating to form a cluster system that a single UAV cannot complete or difficult to complete. In general, the multi-UAV collaborative area coverage path planning is that multiple UAV coordinate with each other to traverse the coverage area to optimize the overall performance. At present, Chen Hai et al. proposed a multi-UAV collaborative area coverage path planning method in "Multi-UAV Collaborative Coverage Path Planning". Wei RuiXuan proposed a control model in the "Full UAV Collaborative Search Fuzzy Cognitive Decision Method", it realizes the cognitive matching and division of tasks, and realizes the search coverage of optimization.

Multi-UAV regional coverage monitoring refers to the regional planning of multiple UAV that the UAV can continuously cover a certain area to achieve long-term continuous detection. Under normal circumstances, each UAV will statically cover some areas. At present, there are few studies on this area, especially for irregular areas.

2 Solution Overview

To solve the multi-UAV regional Coverage Monitoring needs to face two core problems. The first is to divide the coverage area according to the capability of the UAV. The second is the number of UAV in the entire area is as few as possible to meet the coverage requirements. In this paper, the PSO algorithm with branch reduction mechanism is adopted to carry out the UAV regional coverage optimization problem, and with Multi-Agent technology to simulate the particles by the agent in PSO algorithm [3]. Through these two techniques, the particle swarm can automatically perceive whether the initialization scheme is worthy of iteration during optimization. In the iterative process, the branch reduction mechanism is used to tailor the non-optimal particle scheme [4], and to improve the optimization and reduce iteration times. Agent automatically increases the number of particles by self-decision method [5], increases the number of UAV. Thereby, this is a better method to solve this problem by adding a self-decision mechanism and self-trimming capability to the PSO algorithm.

3 Agent Technology

Agent theory is a powerful method for system analysis and design [6]. The concept of Agent comes from the field of distributed artificial intelligence and is a basic term for distributed artificial intelligence [7]. The essence of the agent technology is to study how to make one or more entities to decide and do not disturb the user as much as possible, Agent theory can complete complex tasks commissioned by the user, relying on its own capabilities, using various possible methods and techniques [8]. A single agent should understand the user's intentions as accurately as possible, perceive and adapt to the dynamic environment, and effectively utilize the various information, data, and knowledge that may be utilized in the environment to efficiently accomplish the user's tasks.

A multi-agent is a whole consisting of multiple agents. These agents are coordinated and served each other to accomplish a task. Its own goals and behaviors are not restricted by other agents. It is independent and can resolve conflicts between agents for competition or coordination.

The long-term detection problem of the UAV group in the air can be abstracted as the problem that the multi-agent detects the largest total area in a certain space. Each UAV is modeled by a single agent, with data includes the location of the UAV, the amount of change per search and the effective detection area. A group of UAV is modeled by a multi-agent mode. A multi-UAV agent contains multiple single UAV agents, the best detection area of the group and the best detection location set of the group. The multi-UAV agent has self-awareness that automatically plan the optimal position of each UAV based on the detection area.

4 PSO Algorithm

One of the most well-known and popular optimization algorithms in the field of swarm intelligence is the PSO algorithm [9]. In 1995, the PSO algorithm was a new global optimization evolution algorithm invented by Dr. Eberhard and Dr. Kennedy based on the behavior of nature groups [10]. The algorithm mainly simulates the foraging behavior of the flock, the flight of the flock simulates the entire search space. Each flying bird is abstracted into particles in the search space. Each particle represents a potential solution to the optimization problem [11, 12]. It is generated in the group through particle self-cognition and social information sharing in the group to achieve an optimal solution. The PSO algorithm starts from the random solution and finds the optimal solution through iteration [13, 14]. It finds the global optimal by following the current searched optimal value.

5 Planning Method

5.1 Agent Particle Initialization

Randomly initialize the position and velocity of each particle in the population. The initial position of the particle is randomly generated within the length and width of the equivalent rectangular detection area. The initial velocity of each particle is 0.1 times the length and width of the equivalent rectangular detection area in the UAV detection area. After the particle is generated, the initialization scheme is detected. When the number of particles is greater than 2, if the agent particles have a UAV detection radius that is less than 0.2 times of each other, the random initialization is performed again until the requirement is satisfied.

5.2 Optimization Goal

The optimization problem is to solve the optimal position point by using the detection area ratio of the drone to the detection area as an index function. The detection area ratio is the ratio of the total area detected by the UAV to the area of the area to be covered. Therefore, the problem of optimal allocation of the optimal flight control points $(x_1, y_1), (x_2, y_2) \cdots (x_n, y_n)$ for solving the UAV group is transformed into the optimization process of the objective function, that is, whether the fitness of each particle (the ratio of the detection area obtained from the calculation of the particle point) satisfies the detection index. The requirement is that the detection index is at most 1, that is, full coverage. Generally, in practical applications, in order to reduce the number of UAV, 0.6 to 0.9 is taken.

5.3 The Iterative Process

In general, two extreme values are used to find the optimal solution in PSO algorithm. The first one is the optimal solution found by the particle itself. This solution is called the individual extreme value pBest. The other extreme value is the optimal solution currently found for the entire particle swarm. This extreme value is the global extremum gBest. In this study, the position and fitness of each particle of UAV single agent are stored in the pBest, and the position and fitness of the most adaptive individual in all particles pBest are stored in the gBest of the UAV multi-agent particle.

In each iteration, a single agent updates itself by tracking two "extreme value". The speed and position of the particles update according to the formula and enter the next generation. the speed of the particles update as follows.

$$\begin{aligned} V_{xn}(t+1) &= \omega V_{xn}(t) + c_1 r_{1n}(t)(\overline{x_n} - x_n(t)) + c_2 r_{2n}(t)(\overline{x_n} - x_n(t)) \\ V_{yn}(t+1) &= \omega V_{yn}(t) + c_1 r_{1n}(t)(\overline{y_n} - y_n(t)) + c_2 r_{2n}(t)(\overline{y_n} - y_n(t)) \end{aligned} \tag{1}$$

The particles position update is as follows.

$$\begin{aligned} x_n(t+1) &= x_n(t) + V_{xn}(t+1) \\ y_n(t+1) &= y_n(t) + V_{yn}(t+1) \end{aligned} \tag{2}$$

$(\overline{x_n}, \overline{y_n})$ is the n particle's best position in all the best positions before, $(\overline{x}, \overline{y})$ is the particle's best position in all the best positions of the particle swarm. $(x_n(t), y_n(t))$ is the he position of the n-th particle of the t-th generation, $(v_{xn}(t), v_{yn}(t))$ is the speed in the x-direction and y-direction of the n-th particle of the t-th generation in the solution space. c_1 and c_2 (0.4 ~ 0.7) is the acceleration factor, $\omega(0.6 \sim 0.9)$ is the inertia weight factor.

5.4 Area Calculation

The detection area ratio can be calculated a variety of ways. In order to simplify the calculation of area integral, the target area is divided into M equal parts along each direction, and the grid size is 1 square unit, so the ratio of the calculated area integral problem is approximated by the summation problem.

The coordinate of each cell center node is $(x_j - 0.5, y_j - 0.5)$. The distance from the coordinate of each cell center node to the coordinate of one particle is

$$L = sqrt\left[\left(x_j - 0.5 - x\right)^2 + \left(y_j - 0.5 - y\right)^2\right] \tag{3}$$

By comparing the coordinate distance from the center node of each cell to the coordinate distance of the particle positon L and the UAV's detection radius R, the number of detection cells was determined as *temp*. The detection area ratio is

$$W = \frac{temp}{S_{(MN)}} \tag{4}$$

5.5 Optimization Results

The detection area ratio of each single agent in each generation multi-agent is calculated by

$$p_{tn} = f(x_n(t), y_n(t)) \tag{5}$$

p_{tn} is the detection area ratio of the n-th particle of the t-th generation, $(x_n(t), y_n(t))$ is the position of the n-th particle of the t-th generation, $f(x, y)$ is the calculation function of the detection area ratio. If the detection area ratio of the optimal solution at t-th generation is the gBest of multi-agent, then the iteration is stopped and the particle position of the optimal solution to gBest is output as the flight control point. If not, the iteration is performed and the particle position is updated.

The initial flight control point is determined by increasing the amount of UAV step by step. The detectable area was divided into grids. The average detection area ratio was taken as the objective function, and PSO was used to solve the optimal position of each UAV when the average detection area ratio was the maximum.

6 Pruning PSO Planning Method

PSO algorithm based on pruning can be used to solve optimization problem, the multi-agent technology solve the problem of UAV group for self-perception and decision-making, the UAV group, namely multi-agent particles, can automatically initialize the perception agent particle whether meet the requirements of the iteration, whether to need to be pruned, and decide to optimize the iteration or increase the agent number, thus the PSO algorithm to optimize for automatic perception and decision making process, promote efficiency of optimization.

For the first time, set the number of smart body particles in the multi-agent particles to 1, and then complete the particle initialization and iteration. If the coverage requirements cannot be met, increase the number of particles in the multi-agent particles one by one until the coverage requirements are met.

In each iteration, pregBest is used to record the gbest of all multi-agent particles in the last iteration. If gbest is larger than pregBest, continue to increase the number of particles in the multi-agent particles. Otherwise, in the case of the current number of particles, reinitialize and complete the iteration until the requirement of gbest larger than pregBest is met.

In each iteration, if gbest achieves the largest fitness, namely when the UAV detection area do not overlap and each UAV and UAV's detections are in target detection area of the region, which means that the optimization has reached optimal fitness, stop the iteration, compare detect area ratio with the requirement of detecting area ratio, if does not meet the requirements, directly to increase the amount of multi-agent particles.

In each iteration, if the particles in the multi-agent particles are too close, that is, the detection radius of the particles is less than 0.2 times of the mutual distance of the UAV, the multi-agent particles are deleted from the whole set to achieve pruning operation and ensure the optimization efficiency. In order to guarantee the number of particle swarm, the minimum value of the number of multi-agent particles is set to 20, that is, when the number of multi-agent particle swarm is less than 20, the pruning operation is not carried out.

7 The Simulation Verification

Simulation for 40 km by 80 km of the rectangular area (also can adopt irregular area), using the probe radius of 8 km of UAV for continuous detection, setting detecting area ratio is 0.7, the resulting number of UAV and detection area such as shown in Fig. 1. Red indicates before optimization and blue indicates after optimization, there is no significant difference between the two. Optimization time and detection area such as shown in Fig. 2. Red indicates before optimization, blue indicates after optimization, and the effect is obvious after optimization. The optimized algorithm under the condition of the same number of UAV, obviously improve the optimization time.

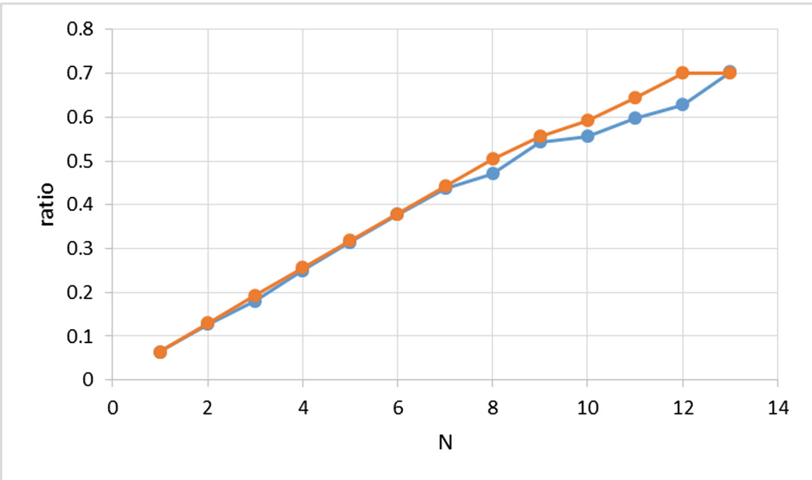


Fig. 1. The resulting number of UAV and detection area.

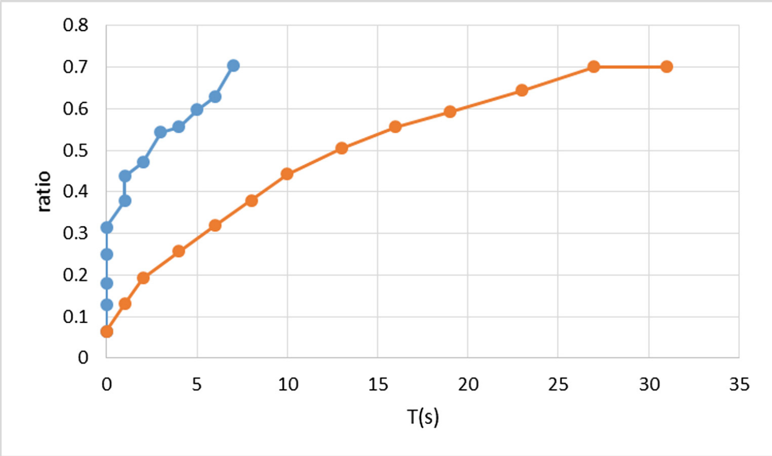


Fig. 2. The optimization time and detection area.

8 Conclusion

The improvement of the detection coverage under multiple UAV formation in a complex environment, affect application of the UAV group in the future.

This article use branch pruning PSO planning method and the combination of multi-agent technology. Multi-agent particle swarm can be used to detect the advantages of the current particle scheme to tailor the bad particles. Particle swarm can automatically increase the number of particles by using multi-agent decision technology, so as to realize the self-decision of the group. In this paper, the computational efficiency of particle swarm optimization is improved, and the UAV coverage optimization problem with a large number of rapid solutions is realized. It has important guiding significance for engineering application.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028.

References

1. Bouvry, P., Chaumette, S., Danoy, G., et al.: Using heterogeneous multilevel swarms of UAVs and high-level data fusion to support situation management in surveillance scenarios
2. Wietfeld, C., Kai, D.: Cognitive Networking for UAV Swarms. Springer, Netherlands (2015)
3. Nasiri, J.A., Fard, A.M., Naghibzadeh, M., et al.: High dimensional problem optimization using distributed multi-agent PSO. In: Third Uksim European Symposium on Computer Modeling and Simulation, pp. 245–250. IEEE Computer Society (2009)

4. Lin, Z., Chen, Z., Fu, M.: A linear control approach to distributed multi-agent formations in d-dimensional space. In: Decision and Control, pp. 6049–6054. IEEE (2013)
5. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Machine Learning Proceedings, pp. 157–163 (1994)
6. Hasegawa, K., Noto, M.: Swarm intelligence algorithm for optimality discovery in distributed constraint optimization. In: IEEE International Conference on Systems, Man and Cybernetics, pp. 3611–3616. IEEE (2014)
7. Olfati-Saber, R., Fax, J.A., Murray, R.M.: Consensus and cooperation in networked multi-agent systems. Proc. IEEE **95**(1), 215–233 (2007)
8. Zhou, Y., Wang, J., Yin, J.: Genetic particle swarm optimization based on multiagent model for combinatorial optimization problem (2008)
9. Afaq, H., Saini, S.: Swarm intelligence based soft computing techniques for the solutions to multiobjective optimization problems. Int. J. Comput. Sci. Issues **8**(3), 498 (2011)
10. Sakti, A.M., Jovan, F., Rys, D., et al.: Modified PSO algorithm for odor source localization problems: progress and challenge. Cloud.snappages.com
11. Frans, V.D.B., Engelbrecht, A.P.: A cooperative approach to particle swarm optimization. IEEE Trans. Evol. Comput. **8**(3), 225–239 (2004)
12. Kumar, V., Minz, S., Kumar, V., et al.: An optimal multi-view ensemble learning for high dimensional data classification using constrained particle swarm optimization. In: Information, Communication and Computing Technology (2017)
13. Chang, F.-C., Huang, H.-C.: A Survey on Intelligent Sensor Network and Its Applications. J. Netw. Intell. **1**(1), 1–15 (2016)
14. Wang, Z., Chen, K., He, L.: AsySIM: modeling asymmetric social influence for rating prediction. Data Sci. Pattern Recognit. **2**(1), 25–40 (2018)

Computational Study the Effects of Venting Pressure and Ignition Location on the Fuel-Air Mixture Explosion Load in Vented Chamber

Xingxing Liang[1(✉)], Zhongqi Wang[1], Pengyi Li[1], and Fuquan Zhang[2,3(✉)]

[1] State Key Laboratory of Explosion Science and Technology, Beijing Institute of Technology, Beijing 100081, China
286943941@qq.com, 491453690@qq.com, czqwang@bit.edu.cn

[2] Digital Performance and Simulation Technology Laboratory, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China
8528750@qq.com

[3] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang Univeristy, Fuzhou 350121, People's Republic of China

Abstract. Numerical analysis were performed to study venting explosion load of methane-air in vented chamber. The effects of venting pressure and ignition location on indoor gas explosion overpressure are discussed. The simulation results showed the expected trends are agreed with the experiments in some extent. Several peaks in overpressure were identified along the pressure–time curves. There are two typical peak overpressures in our study. The first peak is due to the vent failure. The second is caused by the external explosion. It's not a monotonic relationship between the peak overpressure and venting explosion pressure. The influence of the venting pressure on the explosion overpressure is irregular when the venting pressure is less than 10 kPa. Then the explosion overpressure increases with the increase of the venting pressure. The results show that, the maximum peak value of overpressure is shear ignition. But the local overpressure on the pressure relief wall is stronger due to turbulence in central ignition case. Explosion venting process couples the turbulent flow and chemical reaction of combustible medium which is influenced by several parameters.

Keywords: Explosion venting · CFD · Venting pressure · Ignition location

1 Introduction

Explosion venting is a frequently-used way to prevent or lower damage to equipment, pipes or buildings during an accidental explosion. Researchers have been conducted several experiments and numerical analysis to study the blast process during vented explosions. Cooper et al. [1]. conducted central-ignition explosions with natural gas in near-cubic vessels equipped with low-failure pressure blowout panels, to investigate the physical mechanisms responsible for the generation of significant pressure peaks.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 239–246, 2019.
https://doi.org/10.1007/978-3-030-04585-2_29

Van Wingerden [2] also conducted methane explosions in a room and confirmed that external explosions represent a serious blast hazard when the ignition takes place near the rear wall of the room. Explosion venting process couples the turbulent flow and chemical reaction of combustible medium which is influenced by fuel properties, vessel characteristics, venting pressure and size, energy and position of ignition, barrier, etc. [3]. Vented explosion mechanisms have been extensively studied by research institutions and fuel companies. Many different methods have been proposed to design the venting devices for specific locations. Currently, the two most widespread and internationally accepted measures are those recommend in the American standard NFPA 68 [4] and the European standard EN 14797 [5]. However, under certain conditions the two aforementioned criteria can often have incompatible commendation due to the complex nature of the explosion venting process itself. Despite the plethora of researches which have been conducted in recent decades, our understanding of the venting mechanism remains incomplete. Moreover, numerical methods have become pretty powerful tools in various fields of engineering and study. Simulations based on CFD (Computational Fluid Dynamics) are applied to research complex fluid flow problems, containing those including chemical reactions and heat transfer.

The purpose of this paper is to study the influence of ignition location and venting burst pressure on explosion venting in real-scale empty room using commercial CFD program Fluidyn-Ventex.

2 Simulation of Vented Explosion of Methane-Air Mixtures in the Chamber with Variable Vented Pressure and Ignition

2.1 Governing Equation

FLUIDYN-VENTEX uses finite volume method to solve mass, momentum, and energy conservation equation in three-dimensional Cartesian coordinates [6].

Mass conservation equation:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho U) = S_\rho \tag{1}$$

Momentum conservation equation:

$$\frac{\partial (\rho U)}{\partial t} + \nabla \cdot (\rho U) = \nabla \cdot \tau - \nabla p + S_U \tag{2}$$

Energy conservation equation:

$$C_P\left[\frac{\partial (\rho T)}{\partial t} + \nabla \cdot (\rho U T)\right] = -\nabla \cdot q + \left[\frac{\partial p}{\partial t} + U \cdot \nabla p\right] + \tau : \nabla U + S_T \tag{3}$$

Where:

ρ - density, U - velocity vector, T - temperature, p- pressure, $\tau = \mu\dot{\gamma} - (2/3\mu - K)(\nabla \cdot U)\delta$, the viscous stress tensor, $\dot{\gamma} = (\nabla U) + (\nabla U)^T$, the rate-of-strain (or rate-of-deformation) tensor, μ- Effective viscosity, K - Dilatational viscosity (= 0 for a Stokesian fluid), δ - Unit tensor, C_P - specific heat at constant pressure, q - heat flux vector $= -k\nabla T$, k, effective thermal conductivity, S_U - source term for momentum equation (includes body force due to gravitational acceleration), S_T - source term for temperature equation.

Turbulence model is described by $k - \varepsilon$ Model. It solves the transport equations for turbulent kinetic energy, k(Eq. 4), and its dissipation rate, ε (Eq. 5):

$$\frac{\partial(\rho\varepsilon)}{\partial t} + \nabla \cdot (\rho U \varepsilon) = \nabla \cdot \left(\mu_I + \frac{\mu_t}{\sigma_\varepsilon}\right)\nabla\varepsilon + \frac{\varepsilon}{k}\left[C_{\varepsilon 1}(P_k + P_b) - C_{\varepsilon 2}\rho\varepsilon\right] - \left(\frac{2}{3}C_{\varepsilon 1} - C_{\varepsilon 3}\right) \rho\varepsilon\nabla \cdot U + S_\varepsilon^V \tag{4}$$

$$\frac{\partial(\rho k)}{\partial t} + \nabla \cdot (\rho U k) = \nabla \cdot \left(\mu_I + \frac{\mu_t}{\sigma_k}\right)\nabla k + P_k + P_b - \rho\varepsilon - \frac{2}{3}\rho k \nabla \cdot U + S_k^V \tag{5}$$

The present explosion model, called modified BML model, is the modification of the Bray-Moss-Libby model (BML) [8] to include the advantage of the SIF flame model. In BML model a turbulent premixed flame in a highly turbulent flow field is described as a statistical collection of infinite thin laminar flamelets. The SIF is a two fluid model where the flame is represented as an interface between two fluids, reactants and products. The SIF model consists of two parts. In the modified BML flame model, the term k/e used in is replaced by $A_g * U_t$ per unit volume where is turbulent dissipation, k is turbulent kinetic energy, A_g is grid flame surface area and U_t is turbulent burning velocity for the particular fuel–air mixture as below.

$$\omega = C_1 \bar{\rho}\, \Gamma_k A_g U_t (1 + \tau) \frac{\bar{c}(1 - \bar{c})}{(1 + \tau\bar{c})} \tag{6}$$

$$\bar{\rho} = \frac{\rho_R}{(1 + \tau\bar{c})} \tau = \left(\frac{T_b}{T_u} - 1\right) \tag{7}$$

Where:

Γ_k - normalized stretch rate, C_1 - empirical coefficient, τ - expansion ratio, $\bar{c}$ - overall progression variable, T_b - burned product temperature,T_u- unburned reactant temperature, $\bar{\rho}$- normalized density particular fuel–air mixture, The turbulent burning velocity calculated based on the local mixture laminar flame speed, gas viscosity, integral turbulent length scale and the turbulence average fluctuating velocities.

2.2 Calculation Model

In order to compare with W. P. M. Mercxs' experiment, the dimensions of the chamber used in this simulation, 4 m × 3.7 m × 2.6 m, were similar to those in Mercxs' experiment [7]. The chamber was filled by 10% v/v methane-air mixtures. In the experiment the unburned mixture was contained by attaching thin polypropylene sheeting over the vent opening. Polyethylene sheeting has a static failure pressure comparable to, for example, normal glass windows. Therefore, unconstrained vent was arranged and walls are rigid insulation. The geometry of model and monitoring simulation results obtained were compared with the experimental results of P. M. Mercxs et al. There are many monitoring points we have set, only 5 typical monitoring points are selected to analysis in this paper as shown in Fig. 1. Point 1 is on the center of the rear wall, point 4 is on the center of the side wall, point 4 and point 9 is on the wall of the relief surface, and point 22 is the center of the ceiling top. Then ignition position and vented pressure in simulation are listed in Table 1 cases 1–12. The ignition source was effected at three locations, at the rear wall, in the center of the enclosure and near the vent opening. Case 3 is used to validate the model parameter, and the others are designed to study the effect of ignition position and vented pressure indoor vented explosion.

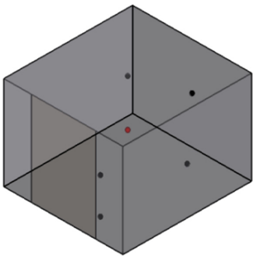


Fig. 1. Physical model of numerical computation

3 Simulation Results and Analysis

3.1 Model Validation

In order to verify the correctness of the model, simulation results obtained were compared with the experimental results of W. P. M. Mercxs et al. Figure 2 shows that the numerical calculation roughly agrees with the experimental results. Both exits two peak overpressure, on is reduce by vent failure, another one is caused by external explosion. The relative errors between the experiment and this simulation are all smaller than 10%. It is seen that the numerical model is valid to reproduce the experimental results.

Table 1. Concentration and vented pressure in simulation

Cases	1	2	3	4	5	6	7	8	9	10	11
Venting pressure (kPa)	0	2	4	6	8	10	20	30	40	10	10
Center	√	√	√	√	√	√	√	√	√	√(Rear wall)	√(Near vent opening)

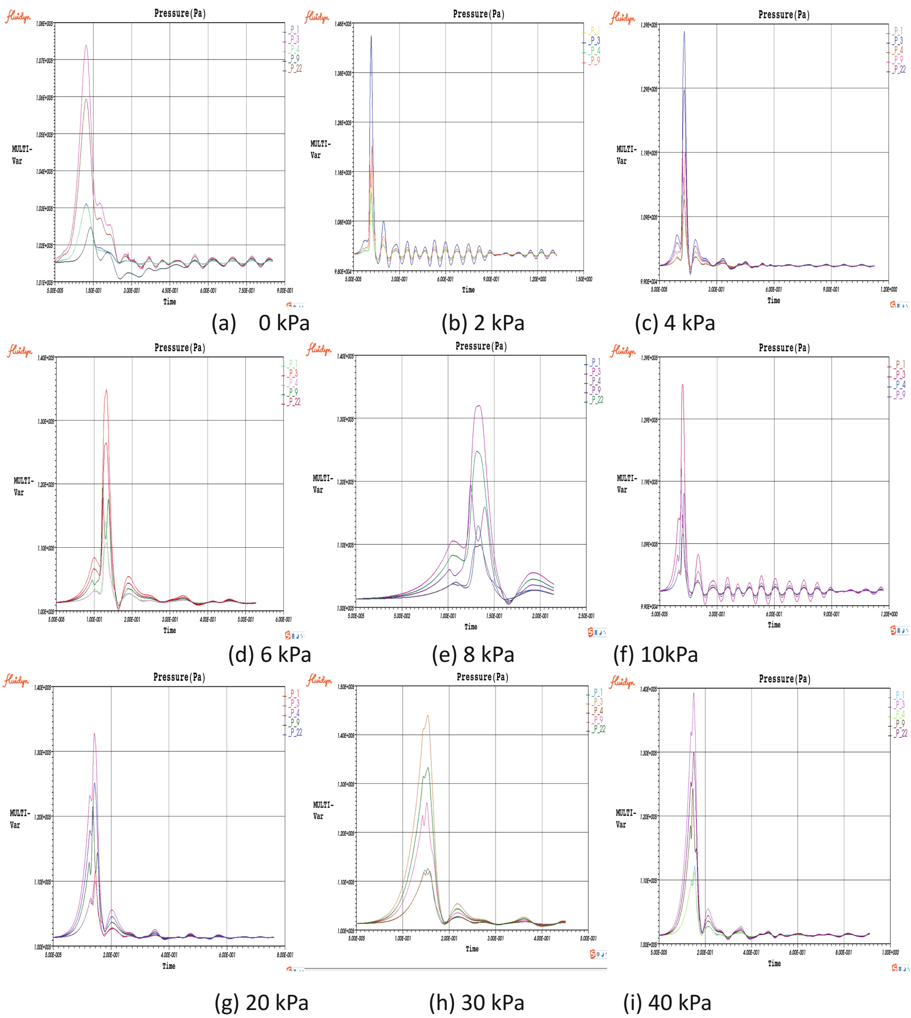


Fig. 2. The pressure-time history of cases 1–9. (Point 1 : the center of the rear wall; point 3 : the center of the side wall, point 4 and point 9 : the wall of the relief surface, and point 22 is the center of the ceiling top.)

3.2 Variation of the Vent Opening Pressure

Vented explosions typically exhibit a series of pressure peaks; of which, not all are present in all explosions [1] G. Tomlin The effect of vent size and congestion in large-scale vented natural gas/air explosions]. There are two typical peak overpressures in cases 2–5 as showed in Fig. 2(b), (c), (d) and (e) respectively. The overpressure is mainly depended on the interplay of two factors. One is the pressure generated by combustion of gas inside the chamber, and another is the gas leakage from the vent which is related to the weak area and the pressure in the enclosure. The first peak is due to the vent failure. The second is caused by the external explosion. A part of unburned mixture is expelled from the chamber after the vent failure in the case of low vented pressure as shown in Fig. 3. When the pressure rise caused by gas explosion in the chamber is not enough to compensate the pressure decline resulted from vent failure, the first peak declines. Subsequently, the flame is drawn toward the vent. The flame area is increased and the chemical reaction rate goes up. At the same time, due to the friction between the vented mixture and the vent, the turbulence near the vent is enhanced. The chemical reaction rate continuously increases. Then the overpressure reaches the second peak following by the unburned mixture explosion. The other cases

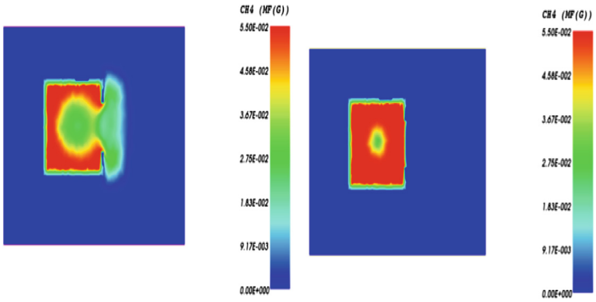


Fig. 3. Mass fraction of CH_4 at first and peak in case 3 are presented in (left) and (right) respectively

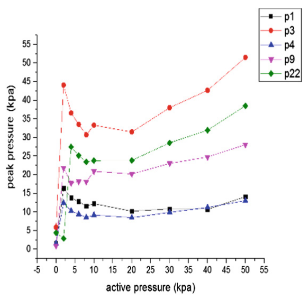


Fig. 4. Maximum pressure peaks variable with venting pressures on several monitoring point

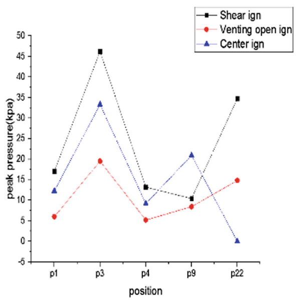
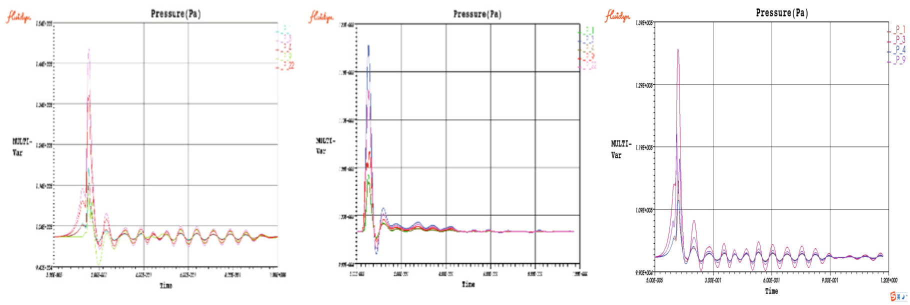


Fig. 5. Maximum pressure peaks variable with monitoring point on several ignition location

are only on typical peak overpressures as showed in Fig. 2(a), (f), (g), (i), (h). Maximum pressure peaks variable with venting pressures on several monitoring point are shown in Fig. 4. It can be seen that the influence of the venting pressure on the explosion overpressure is irregular when the venting pressure is less than 10 kPa. Then the explosion overpressure increases with the increase of the venting pressure.

3.3 Variation of Ignition Source Location

The pressure-time history of cases 6, 10 and 11 are shown in Fig. 5. Maximum pressure peaks variable with monitoring point under different ignition condition such as near rear wall center ignition, near vent open ignition or center ignition are shown in Fig. 6. The results show that, the maximum peak value of overpressure is shear ignition. But the overpressure on the pressure relief wall is stronger due to turbulence in central ignition case.



(a) (b) (c)

Fig. 6. The pressure-time history of cases 6, 10 and 11. ((a): near rear wall center ignition; (b): near vent open ignition; (c): center ignition;)

4 Conclusion

It's not a monotonic relationship between the peak overpressure and venting explosion pressure. The process of venting explosion strongly depends on the various parameters. The maximum peak value of overpressure is shear ignition. But the overpressure on the pressure relief wall is stronger due to turbulence in central ignition case. The influence of the venting pressure on the explosion overpressure is irregular when the venting pressure is less than 10 kPa. Then the explosion overpressure increases with the increase of the venting pressure. Explosion venting process couples the turbulent flow and chemical reaction of combustible medium which is influenced by several parameters. It's not yet fully understood the flame propagation process inside and outside the vent. We should pay attention to the mechanism of explosion venting.

Acknowledgment. The authors would like to acknowledge the financial supports of the National Natural Science Foundation of China (Project No. 51678050). This study was also supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204)

References

1. Cooper, M.G., Fairweather, M., Tite, J.P.: On the mechanisms of pressure generation in vented explosions. Combust. Flame **65**(1), 1–14 (1986)
2. Van Wingerden, K.: Prediction of pressure and flame effects in the direct surroundings of installations protected by dust explosion venting. J. Loss Prev. Process Ind. **6**(4), 241–249 (1993)
3. Shao, H., Jiang, S., et al.: Application and effect of negative pressure chambers on pipeline explosion venting. Loss Prev. **41**, 8–17 (2015)
4. NFPA 68: Standard on Explosion Protection by Deflagration Venting, National Fire Protection Association, Quincy, MA, USA (2013)
5. EN 14797: Explosion venting devices. British Standards Institution, UK (2006)
6. Fluidyn VENTEX-GUI User Manual, June 2017
7. Mercx, W.P.M., van Wingerden, C.J.M., Pasman, H.J.: Venting of gaseous explosions. Process Saf. Prog. **12**(1), 40–46 (1993)

Hybrid Optimization Algorithm of Particle Swarm Optimization with Lagrangian Relaxation for Solving the Multidimensional Knapsack Problem

Jinyan Luo[1(✉)], Geng Lin[1], Fuquan Zhang[2,3], and Lin Xu[4]

[1] Department of Mathematics, Minjiang University, Fuzhou 350108, People's Republic of China
jyluo_jms@sina.com, lingeng413@163.com

[2] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou 350121, People's Republic of China
8528750@qq.com

[3] Digital Performance and Simulation Technology Laboratory, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, People's Republic of China

[4] Key Laboratory of Nondestructive Testing, Fuqing Branch of Fujian Normal University, Fuzhou 350300, People's Republic of China
71471418@qq.com

Abstract. A hybrid algorithm that integrates PSO with Lagrangian relaxation is proposed for solving the multidimensional knapsack problem (MKP). An efficiency measure for MKP based on the LR dual information is defined to combine the object function and the constraints of the MKP together. The efficiency measure is used to determine the core problem for MKP with the aim of reducing the problem scale. Then a hybrid algorithm combines the Quantum Particle Swarm Optimization with a local search method is presented to solve the core problem. Numerical experiments are made on certain knapsack problems and computational results show that the proposed algorithm is very promising.

Keywords: PSO · Lagrangian relaxation · Core problem · MKP

1 Introduction

The MKP is a well known NP-complete combinatorial optimization problem [1]. The MKP arises in several practical problems such as capital budgeting, cargo loading, cutting stock problem and processors allocation in large distributed systems [2]. Recently, using relaxation techniques to extract the intrinsic information of MKP and to guide the search process of intelligent algorithms is a trend to solve the MKP [3, 4]. On the other hand, with the increasing scale of practical problems, the demand for high dimensional MKP algorithm is increasing. In this paper we propose a hybrid approach (LRQPSO-MKP) to solve this problem, which combines the quantum particle swarm

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 247–254, 2019.
https://doi.org/10.1007/978-3-030-04585-2_30

optimization (QPSO) and Lagrange relaxation (LR). Unlike the conventional algorithms, the approach does not directly solve the original MKP. Instead, it firstly defines an efficiency measure for the MKP using LR dual information and then constructs a low-dimensional core problem of the problem, Finally, it uses the QPSO algorithm to solve the problem. And the efficiency measure can be used as an inspired factor of the PSO, guiding the algorithm to focus search in the promising solution area, and improving the search efficiency.

2 The Model of MKP and Its Lagrangian Relaxation

Given a set of items with corresponding values p_i and costs w_{ij}, select some of them to put in the knapsack, without extrapolating their capacities c_i, in such a way that maximizes the sum of values. The MKP can be formulated as follows:

$$(P)\begin{cases} \max f(x) = \sum_{j=1}^{n} p_j x_j \\ subject\ to \sum_{j=1}^{n} w_{ij} x_j \le c_i, i = 1, \ldots, m \\ \quad x_j \in \{0, 1\},\ j = 1, \ldots, n \end{cases} \tag{1}$$

If we multiply the constraints by the penalty factors (i.e. the multipliers) λ_i, and add the resulting expressions to the objective function. Then we can eliminate the original variables and obtain the dual problem of the original MKP as follows:

$$(LD)\begin{cases} \min \Phi(\lambda) = \sum_{i=1}^{m} \lambda_i c_i + \sum_{j=1}^{n} \max\{p_j - \sum_{i=1}^{m} \lambda_i w_{ij}, 0\} \\ s.t.\ \lambda \ge 0 \end{cases} \tag{2}$$

The dual function in Eq. (2) is piecewise linear and the zero point is non-smooth. In general, solving the dual problem need cutting-plane or sub-gradient method [5]. Because the problem of zigzaging is easy to occur in the process of the methods and the performance of the methods is low, we use aggregate function [6], the dual problem can be transformed into a differentiable optimization problems as below:

$$(LD_h)\begin{cases} \min \overline{\Phi}(\lambda) = \sum_{i=1}^{m} \lambda_i c_i + \sum_{j=1}^{n} \{\frac{1}{h} \times \ \ln\{\exp[h(p_j - \sum_{i=1}^{m} \lambda_i w_{ij})] + 1\}\} \\ s.t.\ \lambda \ge 0 \end{cases} \tag{3}$$

where h is the aggregate parameter. Once the multipliers have been found, the Lagrangian relaxation problem can be solved directly and provide a heuristic solution to the original MKP. For this research we define and use the following efficiency measure:

$$\gamma(\lambda)_j = p_j - \sum_{i=1}^{m} \lambda_i w_{ij}, \tag{4}$$

Let $x_L^*(\lambda)$ be the optimal solution to LD-MKP. According to the observation, one can find that x_i^* tends to be 1 in the optimal solution of MKP if the modified profit γ_j takes large positive values, while it tends to be 0 if it takes large negative values.

3 Proposed Strategies and LRQPSO-MKP

3.1 The Core Problem of the MKP

The concept of core problem was first proposed in the literature [7] for the classical one-dimensional 0–1 knapsack problem. The main idea is to reduce the original problem to a core of items for which it is difficult to determine whether or not they will occur in an optimal solution. Puchinger et al. [3] discussed several ways to expand the core problem approach to the MKP. Hill et al. [8] proposed a core problem heuristic approach for the MKP. Here we proposed a method to determine the core problem based on the efficiency measure of the item. The approximate core is identified as follows. Let N be the set of decision variables form (1), $r_{\max} = \max\{\gamma_j; j = 1, 2, \ldots, n\}$ and $r_{\min} = \min\{\gamma_j; j = 1, 2, \ldots, n\}$. Given $\sigma \in (0, 1)$ be the probability of uncertainty, then we define the following sets:

$$S_C = \{j | x_j \in N, \sigma \cdot r_{\min} \le \gamma_j \le \sigma \cdot r_{\max}\} \tag{5}$$

$$S_O = \{j | x_j \in N, j \notin S_C, \gamma_j > 0\} \tag{6}$$

$$S_Z = \{j | x_j \in N, j \notin S_C, j \notin S_O\} \tag{7}$$

S_C is the set of the variables in the core problem, S_O is the set for the non-core variables fixed to one and set S_Z is the set for the non-core variables fixed to zero. The core problem of the MKP is defined as

$$\begin{cases} \max f(x) = \sum\limits_{j \in S_C} p_j x_j \\ \text{subject to} \sum\limits_{j \in S_C} w_{ij} x_j \le c_i - \tilde{w}_i, i = 1, \ldots, m \\ \qquad x_j \in \{0, 1\}, j \in S_C \end{cases} \tag{8}$$

With $\tilde{w}_i = \sum\limits_{j \in S_O} w_{ij}, i = 1, 2, \ldots, m.$

3.2 The Quantum Particle Swarm Optimization

The ability of the BPSO [9] is not ideal for the combinatorial optimization. Inspired by the quantum mechanics and the PSO principle, Yang, Wang, and Jiao proposed a QPSO [10]. A quantum swarm Q at tth iteration was defined as follows:

$$Q(t) = [q_1(t), q_2(t), \ldots, q_s(t)] \ \ with \quad q_i(t) \in [0,1]^n \ \forall i \in s \tag{9}$$

Where $q_i(t)$ represents the probability of the ith particle to be in the "0" state at the tth iteration; s is the particle population size. Then, how to get a discrete particle vector from a quantum particle vector is based on the following rule:

$$x_{id}(t) = \begin{cases} 1 & if \ U(0,1) > q_{id}(t) \\ 0 & otherwise \end{cases} \tag{10}$$

And the evolution of the QPSO can be described as in Yang et al. [10]:

$$Q^{gb}(t) = \alpha \times X^{gb}(t) + \beta \times (e - X^{gb}(t)) \tag{11}$$

$$Q^{b}(t) = \alpha \times X^{b}(t) + \beta \times (e - X^{b}(t)) \tag{12}$$

$$Q(t+1) = c_1 \times Q(t) + c_2 \times Q^{b}(t) + c_3 \times Q^{gb}(t) \tag{13}$$

Where e is a vector of ones. $Q^{gb}(t)$ and $Q^{b}(t)$ correspond to the global best and the local best quantum particle vector, $X^{gb}(t)$ and $X^{b}(t)$ refer to the global best and the local best discrete particle vector, respectively. α and β are control parameters and satisfy $0 \le \alpha, \beta \le 1$ with $\alpha + \beta = 1$. Finally, $0 \le c_1, c_2, c_3 \le 1$ with $c_1 + c_2 + c_3 = 1$ represent the degree of the belief in oneself, local maximum and global maximum respectively.

3.3 A Local Search Strategy for the MKP

In order to improve the quality of the solutions produced by our algorithm, we propose a local search strategy which is launched as an intensification strategy applied to each new best position obtained during the search process. The main idea of local search is to remove an item from the knapsack and put another outside item into the knapsack for every possible pairwise items, Each pairwise element which contains distinct value 1 or 0 is interchanged for a higher profit. Providing that new achieved vector is a feasible solution and has better fitness value than the previous one through swap operation, then the new vector will substitute for old one. This swap operation continues until all pairwise positions are examined. The principle of the efficiency ratio defined as $\tau_j = p_j / \sum_{i=1}^{m} \lambda_i w_{ij}$. Before the starting of strategy, items are sorted by the ascending efficiency ratio and re-numbered. The pseudo-code of the local search strategy is presented as follows.

Algorithm 1 A local search strategy for the MKP

```
Input: Solution vector x of the MKP
Output: Improved solution vector x*
Set x* ← x, isImproved ← true
while isImproved = true do
    R_i = Σ_{j=1}^{n} w_ij x_j,  i = 1,2,...,m;
    isIproved = false;
    for i = 1,2,...,m do //Remove phase
        j ← n;
        while R_i > C_i and j > 0 do
            if x_j = 1, then
                x_j ← 0, R_i > R_i − w_ij;
            end if
            j = j − 1;
        end while
    end for
    for j = 1,2,...,n do//Add phase
        if x_j = 0 and R_i + w_ij ≤ C_i,  i = 1,2,...,m; then
            x_j ← 1, R_i ← R_i + w_ij,  i = 1,2,...,m;
        end if
    end for
    if f(x) > f(x*), then
        x* ← x, isImproved ← true;
    end if
end while
```

For each infeasible solution, the corresponding item with the lowest value is changed from 1 to 0 until the infeasible solution is converted to a feasible one. Therefore the feasibility of the solutions is guaranteed through such approach.

3.4 Description of LRQPSO-MKP Algorithm

The implementation steps of the algorithm are as follows.

- Step1: Set parameters and initialize the vector of multipliers λ^0. Find optimal solution λ^* for the optimization problems (LD_h) as Eq. (3), here we apply the adaptive gradient algorithm (Adagrad) [11] to solve the optimization problem.
- Step2: Calculate efficiency measure of the items by Eq. (4), set the probability of uncertainty σ, and obtain the core problem of the MKP.
- Step3: Initialize $X(t)$ and $Q(t)$ and evaluate $Q(t)$ by the fitness function Eq. (2). The penalty factor is the optimal multipliers λ^* obtained in step1.
- Step4: Evolve $Q(t)$ using Eqs. (11), (12) and (13).
- Step5: Get $X(t)$ from $Q(t)$ by Eq. (10) and evaluate $X(t)$.
- Step6: Apply the local search strategy if a new best position is found.
- Step7: Update the best solution among $X(t)$, if necessary.
- Step8: If termination criteria are met, then output the best solution of the core problem and terminate the algorithm; otherwise, $t = t + 1$ and go to Step4.

Ultimately the best solution of the core problem of the MKP is merged with the trusted object set (that is the S_O and S_Z) of the original problem, and is returned the output as the global optimal solution of the original problem.

4 Experiments and Discussions

In order to assess the effectiveness and feasibility of the proposed method, Our approach is tested on two sets of benchmark instances. The first set is available in the ORLIB Library [12], and the second set consists of 11 problems with $m = 15$ to 100 and $n = 100$ to 2500. The algorithm is coded in Visual C and run on person computer with a 2.5 GHz Intel Core i5 processor. Except for the special explanation, in the experiment, each instance was tested 25 times independently, and the maximum number of iterations allowed for each test was 2000. In our implementations, we set the learning rate $\eta = 0.25$, the aggregate parameter $h = 10^6$ and the initial multipliers $\lambda^0 = 10e_m$ (e_m is an m-dimensional column vector with element 1). And we set $\alpha = 0.1$ and $\beta = 0.9$, $c_1 = 0.4$, $c_2 = 0.2$ and $c_3 = 0.4$. The value of uncertainty σ is crucial to the performance of the approach, we refer to the literature [8] setting $\sigma = 0.15$.

The LRQPSO-MKP algorithm is compared with two existing algorithms. One is the QPSO* [12] algorithm, another is the TE-BDS [13]. Tables 1 and 2 show the results of the LRQPSO-MKP compared with QPSO* and TE-BDS respectively. The results of the two existing algorithms are directly extracted from the corresponding literature. For the LRQPSO-MKP approach, in addition to the performance of the corresponding solution, we also provide the dimension of the core problem (Dim.), the average time to get the optimal solution at the first time (T-f) and the average time (ACT) consumed after all iterations.

Table 1 provides a summary of the results obtained by LRQPSO-MKP and QPSO* for a reduced set of 18 instances selected among the 270 benchmarks of OR-Library. This set of instances are known to be very hard to solve using exact methods. The LRQPSO-MKP obtained the optimal solution of 13 of these examples in each test. In terms of average performance, LRQPSO-MKP outperforms the QPSO* in 8 instances, and 10 instances are similar each other. Although the hardware platform of the two algorithms is slightly different, but the great difference of the completion times can show that the efficiency of the LRQPSO-MKP is better than the QPSO*. The QPSO* combined a heuristic repair operator instead of the penalty function technique which usually used to avoid the violation of problem constraints. The repair operator consists of two iterative processes, which requires a lot of computation times, especially for high-dimensional problems. Moreover, the first time is much less than the average time, thus it can be seen that the total iteration number needed by LRQPSO-MKP to obtain the same solution quality can be further reduced.

As shown in Table 2, for 11 tested instances, in terms of solutions quality LRQPSO-MKP achieved better results (boldface) than TE-BDS. In addition to mk-gk500.50-08 and mgk1500.50-10, the corresponding variance is all less than the TE-BDS algorithm; the standard variance of LRQPSO-MKP is more consistent in different scale problems; these results show that the LRQPSO-MKP algorithm has strong robustness. In terms of average processing times, the results also show that our method rivals with TE-BDS in the most of instances.

Table 1. Comparison of LRQPSO-MKP with QPSO*.

Prob.		QPSO*			LRQPSO-MKP				
Inst.	Opt.	Best	Mean	ACT (s)	Dim.	Best	Mean	T-f (s)	ACT (s)
mkp100.5-00	24 381	24 381	24 381.0	3.198	27 × 5	24 381	24 381.0	0.401	**1.966**
mkp100.5-20	59 822	59 822	59 822.0	—	29 × 5	59 822	59 822.0	0.592	3.217
mkp100.10-00	23 064	23 064	23 064.0	5.273	31 × 10	23 064	23 064.0	0.653	**3.458**
mkp100.10-20	57 375	57 375	57 375.0	—	36 × 10	57 375	57 375.0	0.656	4.213
mkp100.30-00	21 946	21 946	21 946.0	6.333	41 × 30	21 946	21 946.0	0.413	**4.980**
mkp100.30-20	57 494	57 494	57 494.0	—	41 × 30	57 494	57 494.0	0.926	4.712
mkp250.5-00	59 312	59 312	**59 312.0**	22.232	64 × 5	59 312	59 302.9	1.437	**6.879**
mkp250.5-20	149 665	149 665	149 650.5	—	69 × 5	149 665	149 650.5	1.617	7. 096
mkp250.10-00	59 187	59 182	59 173.0	28.714	71 × 10	59 180	59 171.3	1.578	**7.949**
mkp250.10-20	151 809	151 779	151 769.0	—	75 × 10	151 779	151 769.0	1.772	8.018
mkp250.30-00	56 842	56 796	56 745.5	49.612	86 × 30	**56 842**	**56 791.5**	1.835	**9.136**
mkp250.30-20	150 163	150 096	150 052.0	—	87 × 30	**150 163**	**150 119.0**	1.453	9.216
mkp500.5-00	120 148	120 130	120 105.7	94.396	113 × 5	**120 148**	**120 148.0**	**1.556**	**11.461**
mkp500.5-20	295 828	295 828	**295 797.7**	—	123 × 5	295 828	295 781.5	1.367	12.874
mkp500.10-00	117 821	117 744	117 733.5	109.935	152 × 10	**117 811**	**117 809.0**	3.177	**14.416**
mkp500.10-20	304 387	304 344	**304 329.5**	—	123 × 10	**304 387**	304 307.3	6.385	12.887
mkp500.30-00	116 056	115 991	115 906.0	125.253	166 × 30	**116 014**	**115 919.8**	7.124	**14.578**
mkp500.30-20	301 675	301 643	301 635.0	—	157 × 30	**301 667**	**301 658.9**	12.509	14.793

Table 2. Comparison of LRQPSO-MKP with TE-BDS.

Prob.		TE-BDS			LRQPSO-MKP			
Inst.	Opt.	Mean	Std	ACT (s)	Dim.	Mean	Std	ACT (s)
mk-gk100.15-01	3766	3720.86	7.01	22.5922	38 × 15	**3766.00**	**1.77**	**7.409**
mk-gk100.25-02	3958	3905.62	8.20	24.0208	45 × 25	**3958.00**	**5.20**	**7.606**
mk-gk150.25-03	5650	5542.22	9.87	26.9987	82 × 25	**5579.10**	**7.09**	**15.584**
mk-gk150.50-04	5764	5648.32	**8.50**	31.1158	78 × 50	**5677.48**	8.61	**17.634**
mk-gk200.25-05	7557	7376.84	9.56	29.9581	98 × 25	**7292.00**	**9.07**	**17.091**
mk-gk200.50-06	7672	7504.88	9.34	34.3660	84 × 50	**7586.06**	**8.59**	**19.411**
mk-gk500.25-07	19 215	18600.36	21.70	44.1570	306 × 25	**19075.23**	**13.97**	**35.734**
mk-gk500.50-08	18 801	18308.58	**13.09**	50.0867	324 × 50	**18354.53**	15.47	**43.277**
mk-gk1500.25-09	58 085	56058.74	36.00	**103.5136**	944 × 25	**57826.90**	**25.81**	150.701
mk-gk1500.50-10	57 292	55746.32	30.97	**112.9763**	956 × 50	55276.03	**26.07**	174.445
mk-gk2500.100-11	95 231	93192.98	30.19	224.0877	1520 × 100	**93300.61**	**29.93**	**215.640**

5 Conclusion

According to the Lagrange relaxation dual information, we primarily presented an efficiency measure of the MKP, then based on it the definition of the core problem is provided. The introduction of the core problem greatly reduces the size of the problem, thus reducing the search space of the particle swarm. The proposed efficiency measure can be used indirectly as the heuristic factor of the particle swarm, which guides the algorithm to enhance the search in the potential solution region. In addition, a local search strategy is introduced into our algorithm to improve the quality of solutions. By testing in generous medium and large-scale instances, the advantages of the proposed algorithm in solving the quality and efficiency of the solution are verified.

Acknowledgements. This study was supported by the Research Program Foundation of Minjiang University under Grants No. MYK17021 and supported by the Major Project of Sichuan Province Key Laboratory of Digital Media Art under Grants No. 17DMAKL01 and supported by Fujian Province Guiding Project under Grants No. 2018H0028 and supported by National Nature Science Foundation of China (Grant number: 61871204). We also acknowledge the solution from National Natural Science Foundation of China (61772254), Key Project of College Youth Natural Science Foundation of Fujian Province (JZ160467), Fujian Provincial Leading Project (2017H0030), Fuzhou Science and Technology Planning Project (2016-S-116), Program for New Century Excellent Talents in Fujian Province University (NCETFJ) and Program for Young Scholars in Minjiang University (Mjqn201601).

References

1. Chu, P.C., Beasley, J.E.: A genetic algorithm for the multidimensional knapsack problem. J. Heuristics **4**(1), 63–86 (1998)
2. Hanafi, S., Wulbaut, C.: Scatter search for the 0–1 multidimensional knapsack problem. J. Math. Model. Algorithms **7**(2), 143–159 (2008)
3. Puchinger, J., Raidl, G.R., Pferschy, U.: The multidimensional knapsack problem: structure and algorithms. INFORMS J. Comput. **22**(2), 250–265 (2010)
4. Martins, J.P., Fonseca, C.M., Delbem, A.C.B.: On the performance of linkage-tree genetic algorithms for the multidimensional knapsack problem. Neurocomputing **146**(1), 17–29 (2014)
5. Frangioni, A.: About Lagrangian methods in integer optimization. Ann. Oper. Res. **139**(1), 163–193 (2005)
6. Li, X.-S.: An efficient approach to a class of non-smooth optimization problems. Sci. China (Ser. A) **37**(3), 323–330 (1994)
7. Balas, E., Zemel, E.: An algorithm for large zero-one knapsack problems. Oper. Res. **28**(5), 1130–1154 (1980)
8. Hill, R.R., Kun Cho, Y., Moore, J.T.: Problem reduction heuristic for the 0–1 multidimensional knapsack problem. Comput. Oper. Res. **39**(1), 19–26 (2012). https://doi.org/10.1016/j.cor.2010.06.009
9. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: Proceedings of IEEE International Conference on Computational Cybernetics and Simulation, Orlando, USA, pp. 4104–4109 (1997)
10. Yang, S., Wang, M., Jiao, L.: A quantum particle swarm optimization. In: Proceedings of IEEE Congress on Evolutionary Computation, vol. 1, pp. 320–324, June 2004
11. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011). http://jmlr.org/papers/v12/duchi11a.html
12. Haddar, B., Khemakhem, M., Hanafi, S., et al.: A hybrid quantum particle swarm optimization for the multidimensional knapsack problem. Eng. Appl. Artif. Intell. **55**(C), 1–13 (2016)
13. Kong, X., Gao, L., Ouyang, H., et al.: Solving large-scale multidimensional knapsack problems with a new binary harmony search algorithm. Comput. Oper. Res. **63**, 7–22 (2015)

Tongue Body Localization Based on Image Clustering and Gray Projection

Weixia Liu[1,2,3](✉)

[1] Straits Institute, Minjiang University, Fuzhou, China
Liuweixia0201@126.com
[2] Internet Innovation Research Center, The Fujian College's Research Base of Humanities and Social Science, Minjiang University, Fuzhou, China
[3] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China

Abstract. Tongue diagnosis is often used in Traditional Chinese Medicine (TCM). Tongue image segmentation, which extracts tongue body, is a key step when manufacturing an automated system of tongue diagnosis. Localization of tongue body depicted by a rectangle is a useful preprocessing step in tongue image segmentation, which can eliminate the adverse effect of strong edges from neighboring tissues such as face and lip when extracting tongue body contour. After exploring the existing tongue body localization method based on gray projection, we propose an upgraded method combining image clustering with gray projection. Specifically, our proposed method first conducts the clustering on the image hue component in HSI (i.e., hue, saturation, and intensity) space to determine three thresholds. Then, image thresholding and morphological operations are sequentially performed to generate a binary image, and its largest object region is taken as the initial localization result. Finally, the localization result is refined by performing gray projection on the image red component. Experiments on a variety of tongue images showed that our proposed method significantly improves the accuracy of tongue body localization in comparison with the existing gray projection method.

Keywords: HSI · Image thresholding · Gray projection
Tongue body localization

1 Introduction

Tongue diagnosis [1–3] is often used in TCM due to the virtues such as effectiveness, painlessness, and lack of side-effects. It has a history over 3000 years in China. Practitioners of Chinese medicine have accumulated very rich tongue diagnosis experiences on judging human health status by the tongue body features, such as the color, coating, and texture. Eight classic tongue diagnosis principles [1, 2] deem that different parts of tongue body depict the health status of different human organs. For instance, the appearance of tongue body is a useful indicator of monitoring improvement or deterioration in human health. In the opinion of TCM, syndrome can reflect many diseases and human health status [3].

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 255–265, 2019.
https://doi.org/10.1007/978-3-030-04585-2_31

Modern Western Medicine has increasingly recognized that the human tongue can provide useful clues to the human health status. Accordingly, some researchers [4] in Western Medicine have also used tongue diagnosis to help clinical decision making. For instance, the tongue body coating was found to be strongly related with many viable salivary bacteria, being taken as a risk indicator for aspiration pneumonia in edentate patients [4]. In addition, amyloidosis of the tongue was used to judge plasmacytoma. An increasing number of studies explored the capability of tongue diagnosis to help clinical decision making.

Conventional tongue diagnosis is qualitative, and its diagnosis accuracy is instable due to its high dependence on the experience of the practitioner. Today, powerful computer hardware has made it possible to manufacture automated tongue diagnosis systems via advanced digital image processing [5–7] and pattern recognition techniques [8–10]. These automated systems usually first use image segmentation techniques to extract the tongue body in an image, then employ feature extraction techniques to calculate the tongue body features, and finally use a classifier to achieve a final tongue diagnosis. Therefore, tongue image segmentation, which extracts the tongue body, is a key step in the above process. Several researches [1–3] have been performed to resolve the issue of tongue image segmentation. Unfortunately, tongue image segmentation still faces some challenges caused by large personal variation in tongue body characteristics.

Image segmentation techniques can be categorized into three types, i.e., boundary-based methods [11, 12], region-based methods [13, 14], and hybrid methods [15, 16]. Hybrid methods usually achieve better image segmentation results due to simultaneous use of boundary and region information. Due to the challenges of tongue image segmentation, single image processing technique usually fails to obtain satisfactory segmentation result, some efforts [1–3] have been put on hybrid tongue image segmentation techniques during the past several decades.

To improve accuracy of tongue image segmentation, Zhang and Qin [2] proposed an image preprocessing method. The proposed method uses a gray projection technique to achieve tongue body localization in a tongue image. The tongue body localization tries to seek the rectangle area located by the tongue body. Tongue body localization is helpful for subsequent segmentation of the tongue body, as it can alleviate adverse effect of strong image edges from neighboring tissues such as face and lip on the tongue body contour extraction. After exploring the limitation of the existing gray projection (GP) based localization method [2], we propose an upgraded tongue body localization method combining image clustering with gray projection. Experiments on a variety of tongue images with large personal variation validate the superiority of our proposed method over the existing method.

2 Tongue Body Localization Based on Gray Projection (GP)

The tongue body localization method based on gray projection (GP) [2] was presented according to two prior knowledge. The first knowledge is that tongue body root region is usually darker than other regions on the human face when generating tongue images. The second knowledge is that there may exist dark regions near tongue body contour due to blocked light by stretched tongue body when generating tongue images. These dark regions cause those involved image rows and image columns have lower average gray values than other rows and columns, respectively.

For a given tongue image I with M rows and N columns, let $I(i, j)$ denote the gray level of pixel $p_{i,\ j}$ at the i-th row and j-th column. The detailed process of the gray projection based tongue body localization includes the following four steps.

(1) Calculate average gray value of each image row, where the average gray value of i-th row is formulated as

$$G_r(i) = \frac{1}{N}\sum\nolimits_{j=1}^{N} I(i,j). \tag{1}$$

Similarly, the average gray value of j-th column is formulated as

$$G_c(j) = \frac{1}{M}\sum\nolimits_{i=1}^{M} I(i,j). \tag{2}$$

(2) Find the row with the lowest G_r value among the front half of image rows, which can be formulated as

$$R_{up} = \min\nolimits_{1 \le i \le M/2} G_r(i). \tag{3}$$

Similarly, the row with the lowest G_r value among the latter half of image rows can be determined as

$$R_{down} = \min\nolimits_{M/2 < i \le M} G_r(i). \tag{4}$$

(3) Find the column with the lowest G_c value among the front half of image columns, which can be formulated as

$$C_{left} = \min\nolimits_{1 \le j \le N/2} G_c(j). \tag{5}$$

Similarly, the column with the lowest G_c value among the latter half of image columns can be determined as

$$C_{right} = \min\nolimits_{N/2 < j \le N} G_c(j). \tag{6}$$

(4) Use R_{up}, R_{down}, C_{left}, and C_{right} to confine a rectangle region as the localization result of the tongue body.

After exploring the principle of the GP method on tongue body localization, we observed that the GP method usually cannot achieve satisfactory tongue body localization on acquired original tongue images, but usually achieves satisfactory localization on clipped tongue images. Figure 1 gives an example to validate our observation, where Fig. 1(b) and (d) use blue rectangles to exhibit the localization results achieved by applying the GP method on an original tongue image in Fig. 1(a) and its clipped image in Fig. 1(c), respectively. From Fig. 1, one can conclude that the

GP method obtains good tongue body localization result on the clipped tongue image, but obtains bad localization result on the original tongue image. This demonstrates the limitation of the GP method on tongue body localization.

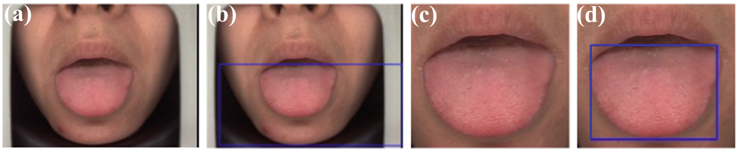


Fig. 1. Tongue body localization results obtained by the GP method [2]. (a) original image, (b) the localization result of (a), (c) the clipped image of (a), (d) the localization result of (c).

3 The Proposed Method

After exploring the gray projection (GP) method [2], we observed that the method is invalid for acquired original tongue images as illustrated in Fig. 1(a). To obtain satisfactory tongue body localization results, the GP method need clip original tongue image, which is time-consuming and inconvenient. To resolve this issue, we propose an upgraded tongue body localization method via image clustering and gray projection. Our proposed tongue body localization method includes the following seven steps.

(1) Color space mapping: an image in RGB color space is mapped into HSI color space via the following equations:

$$H = \begin{cases} \theta, & G \geq B \\ \theta - 2\pi, & G < B \end{cases} \tag{7}$$

$$S = 1 - \frac{3}{R+G+B}\min\{R, G, B\}, \tag{8}$$

$$I = \frac{1}{3}(R+G+B), \tag{9}$$

where

$$\theta = \arccos\left\{ \frac{[(R-G)+(R-B)]/2}{\left[(R-G)^2 + (R-B)(G-B)\right]^{1/2}} \right\}. \tag{10}$$

In Eqs. (7)–(10), R, G, and B are the red, green, and blue components, respectively.

(2) Clustering-based threshold determination: an image clustering scheme is proposed to seek three thresholds for subsequent image thresholding. In this process, only the hue component of an image is used to perform image clustering, the class number is empirically set to 3, and three initial class centers are set to the minimum value, the median value, and the maximum value in the hue component. Three class

centers are updated iteratively until they are unchanged. At each iteration, each pixel's hue value is classified into one of three classes according to the minimum absolute hue difference, and then each class center is updated as its average hue value. After the image clustering process, three final class centers are gathered into a set called T_{set} with ascending order.

(3) Image thresholding and morphological operations: the elements in T_{set} are first used to conduct image thresholding on the image hue component, where the image thresholding result B_{img} is defined as

$$B_{img}(i,j) = \begin{cases} 1, & \text{if } H(i,j) < T_{set}(1) \quad \text{or} \quad H(i,j) > T_{set}(3) \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

where $T_{set}(i)$ denotes the i-th element in T_{set}. After obtaining the thresholding result, i.e., B_{img}, morphological operations (i.e., opening operation and image filling operation) are sequentially performed to obtain the refined result, i.e., RB_{img}, where the morphological opening operation used the structure element with disk shape and radius "4". Figure 2(b) and (c) exhibit B_{img} and RB_{img}, respectively.

(4) Determination of the initial tongue body region: the largest object region (i.e., white region) in the refined image thresholding result, i.e., RB_{img}, is taken as the initial tongue body region and solely remained in a binary image called O_{img} shown in Fig. 2(d).

(5) Row localization based on gray projection: first, row-clipping is performed by applying gray projection to those image rows of the red component containing object pixels in O_{img}. Specifically, average gray value of object pixels on each row containing object pixels is first calculated, and then the row with the lowest average gray value among those image rows containing object pixels is selected as the upper bound, i.e., R_{up}, of the rectangle region located by the tongue body. Second, the maximum row number of the object region in O_{img} is taken as the lower bound, i.e., R_{down}, of the rectangle region located by the tongue body. Since we think that the tongue body should be located between the R_{up}-th row and the R_{down}-th row, the object pixels on those rows out of the row range [R_{up} R_{down}] should be faked tongue body pixels and clipped. The clipped result of Fig. 2(d) is exhibited in Fig. 2(e).

(6) Column localization: the left bound and the right bound of the object region in Fig. 2(e) are taken as the left bound, i.e., C_{left}, and the right bound, i.e., C_{right}, of the rectangle region located by the tongue body, respectively.

(7) Final tongue body localization: the tongue body region is localized by using the row bounds (i.e., R_{up} and R_{down}) and the column bounds (i.e., C_{left} and C_{right}). A green rectangle is used to intuitively exhibit the tongue body localization result obtained by our proposed method as shown in Fig. 2(f).

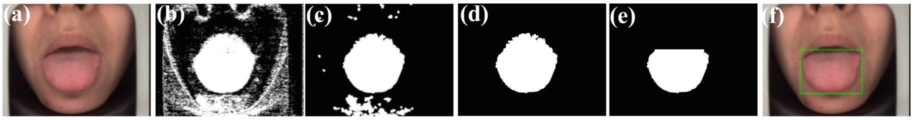


Fig. 2. Visual intermediate results generated by the proposed method. (a) original image, (b) image thresholding result on the hue component of (a), (c) refined image thresholding result after morphological operations, (d) the result only remaining the largest object region of (c), (e) the refined result after row-clipping on (d), (f) final tongue body localization result.

4 Experimental Results

To validate the effectiveness of our proposed method for tongue body localization, extensive experiments were performed on four groups of tongue images with large differences on tongue body size, shape, color, coating and texture. Each image has size of 640×480. Tongue body localization results obtained by our proposed method were compared with those achieved by the gray projection (GP) method [2]. To quantitatively measure the accuracy of tongue body localization, the rectangle region obtained by each tongue body localization method is taken as the object region in a binary segmentation result. Thus, misclassification error (ME) [17] and kappa index (KI) [18] can be used to quantitatively measure the accuracy of tongue body localization. The definitions of ME and KI are as

$$\mathrm{ME} = 1 - \frac{|B_m \cap B_a| + |F_m \cap F_a|}{|B_m| + |F_m|}, \tag{12}$$

$$\mathrm{KI} = 2\frac{|F_m \cap F_a|}{|F_m| + |F_a|}, \tag{13}$$

where B_m and F_m are background and foreground of a manual ideal segmentation result (ground truth), respectively; B_a and F_a are background and foreground of a binarization result corresponding to the tongue body localization result, and $|\cdot|$ is the cardinality of a set. ME and KI range between 0 and 1. Lower ME values indicate higher accuracy of tongue body localization, while higher KI values indicate higher accuracy of tongue body localization.

4.1 Qualitative Comparison of Localization Results

Six original tongue images used in the first group of experiments are shown in the first row of Fig. 3, where the front three images have small tongue bodies, meanwhile the latter three images have large tongue bodies. The manual ideal segmentation results of the six tongue images are listed in the second row of Fig. 3, where the manual ideal tongue body contours are depicted by cyan curves. The localization results obtained by the GP method [2] are shown in the third row of Fig. 3, where the localization results are depicted by blue rectangles. The localization results yielded by our proposed method are shown in the last row of Fig. 3, where the localization results are depicted by green rectangles. From Fig. 3, it can be observed that the GP method can accurately find upper bounds of tongue body roots in most cases, but cannot correctly find lower bounds, left bounds, and right bounds of the tongue bodies. Conversely, the proposed method successfully finds four bounds of the tongue bodies and obtains satisfactory tongue body localization results.

Fig. 3. Localization results of tongue images with large difference on tongue body size from up to down: original images, ground truths, localization results obtained by the gray projection (GP) method [2], localization results yielded by the proposed method.

In addition, Fig. 4 exhibits visual tongue body localization results on six tongue images with large difference on tongue body shape, where the front two images have square tongue bodies, the middle two images have flat tongue bodies, and the last two images have vertical tongue bodies. Figure 5 exhibits visual tongue body localization results on six tongue images with large difference on tongue body color, where the front three images have tongue bodies with slightly white color, and the latter three images have tongue bodies with deep red color. Figure 6 exhibits visual tongue body localization results on six tongue images with strong tongue coating or texture, where the front three images have strong tongue coating, and the latter three images have strong texture. From Figs. 4, 5 and 6, it can be seen that the proposed method successfully achieves tongue body localization on the tongue images with large differences on tongue body shape, color, coating, and texture. However, the gray projection (GP) method usually only finds the locations of tongue body roots, and fails to successfully achieve tongue body localization.

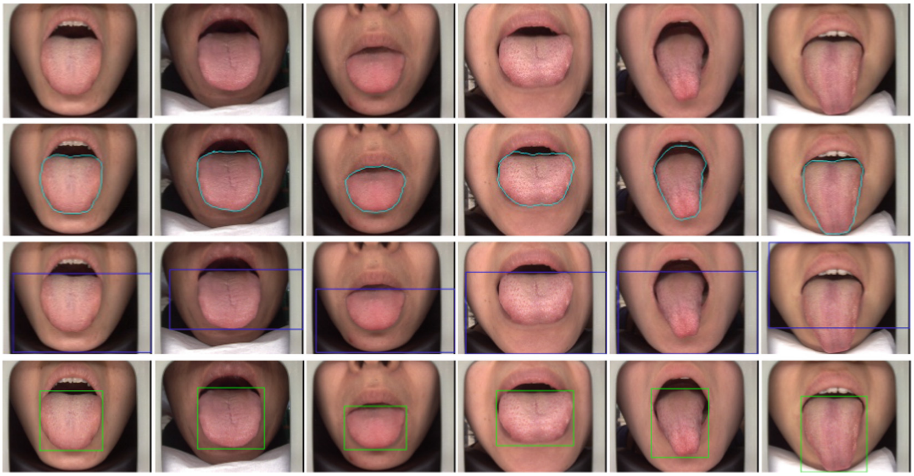


Fig. 4. Localization results of tongue images with large difference on tongue body shape from up to down: original images, ground truths, localization results obtained by the gray projection (GP) method [2], localization results yielded by the proposed method.

Fig. 5. Localization results of tongue images with large difference on tongue body color from up to down: original images, ground truths, localization results obtained by the gray projection (GP) method [2], localization results yielded by the proposed method.

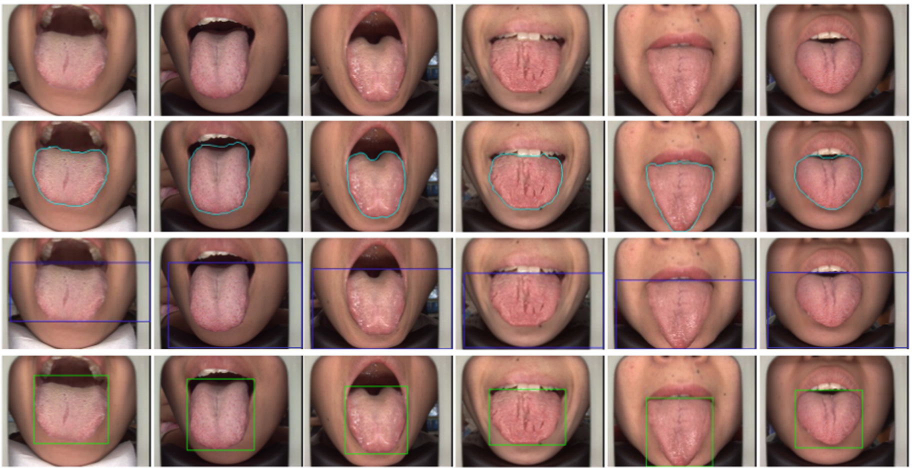


Fig. 6. Localization results of tongue images with strong tongue body coating or texture from up to down: original images, ground truths, localization results obtained by the gray projection (GP) method [2], localization results yielded by the proposed method.

4.2 Quantitative Comparison of Localization Results

After performing the above qualitative comparison on tongue body localization results, we further conduct quantitative comparison on tongue body localization results. We used misclassification error (ME) [17] and kappa index (KI) [18] to quantitatively evaluate those tongue body localization results. The average ME and KI values of tongue body localization results obtained by both methods in four groups of experiments are listed in Table 1. From the table, one can conclude that the proposed method significantly decreases the ME values and increases the KI values in comparison with the gray projection (GP) method. Both ME values and KI values demonstrate that the proposed method has higher tongue body localization accuracy than the gray projection method.

Table 1. Quantitative comparison of tongue body localization results.

Group name	Average ME values		Average KI values	
	GP [2]	Proposed	GP [2]	Proposed
Group 1	0.486	**0.068**	0.427	**0.836**
Group 2	0.473	**0.063**	0.424	**0.852**
Group 3	0.378	**0.063**	0.474	**0.850**
Group 4	0.423	**0.071**	0.489	**0.850**

5 Conclusion

Tongue image segmentation is a key step when manufacturing a computer-aided tongue diagnosis system. Tongue body localization is a useful preprocessing step in tongue image segmentation. The existing gray projection method fails to obtain satisfactory tongue body localization results on original tongue images acquired by camera. To resolve this issue, we presented an upgraded tongue body localization method combining image clustering with gray projection. Specifically, the proposed method first conducts image clustering on the hue component of an image in HSI color space to determine three thresholds. Then, image thresholding and morphological operations are sequentially performed to generate a binary image, and the largest object region in the binary image is taken as an initial tongue body localization result. Finally, the initial localization result is refined by using the gray projection technique on the red component of the image. Experiments on four groups of tongue images with large personal variation demonstrate significant improvement of the proposed method on tongue body localization over the existing method.

Acknowledgment. This work is partially supported by National Natural Science Foundation of China (61772254 and 61202318), Fuzhou Science and Technology Project (2018-S-123 and 2016-S-116), Fujian Provincial Leading Project (2017H0030), and Technology Project of Education Department of Fujian Province (JA15425 and JAT160391).

References

1. Shi, M., Li, G., Li, F., Xu, C.: Computerized tongue image segmentation via the double geo-vector flow. Chin. Med. **9**(1), 7–16 (2014)
2. Zhang, L., Qin, J.: Tongue-image segmentation based on gray projection and threshold-adaptive method. J. Clin. Rehabil. Tissue Eng. Res. **14**(9), 1638–1641 (2010)
3. Wu, K., Zhang, D.: Robust tongue segmentation by fusing region-based and edge-based approaches. Expert Syst. Appl. **42**(21), 8027–8038 (2015)
4. Abe, S., Ishihara, K., Adachi, M., Okuda, K.: Tongue-coating as risk indicator for aspiration pneumonia in edentate elderly. Arch. Gerontol. Geriatr. **47**(2), 267–275 (2008)
5. Xiao, B., Tang, H., Jiang, Y., Li, W., Wang, G.: Brightness and contrast controllable image enhancement based on histogram specification. Neurocomputing **(275)**, 2798–2809 (2018)
6. Wang, X., Chen, L.: An effective histogram modification scheme for image contrast enhancement. Signal Process. Image Commun. **58**, 187–198 (2017)
7. Zhuang, P., Fu, X., Huang, Y., Ding, X.: Image enhancement using divide-and-conquer strategy. J. Vis. Commun. Image Represent. **45**, 137–146 (2017)
8. Chen, X., Cai, Y., Chen, L., Li, Z.: Discriminant feature extraction for image recognition using complete robust maximum margin criterion. Mach. Vis. Appl. **26**, 857–870 (2015)
9. Luo, Y.T., Zhao, L.Y., Zhang, B., Jia, W., Xu, B.Q.: Local line directional pattern for palmprint recognition. Pattern Recogn. **50**, 26–44 (2016)
10. Liu, G., Yang, J., Li, Z.: Content-based image retrieval using computational visual attention model. Pattern Recogn. **48**(8), 2554–2566 (2015)
11. Wang, L., Zou, L., Zhao, S.: Edge detection based on subpixel-speckle-shifting ghost imaging. Opt. Commun. **407**, 181–185 (2018)

12. Rajathilagam, B., Rangarajan, M.: Edge detection using G-lets based on matrix factorization by group representations. Pattern Recogn. **67**, 1–15 (2017)
13. Chondro, P., Yao, C.Y., Ruan, S.J., Chien, L.C.: Low order adaptive region growing for lung segmentation on plain chest radiographs. Neurocomputing **(275)**, 1002–1011 (2018)
14. Hettiarachchi, R., Peters, J.F.: Voronoï region-based adaptive unsupervised color image segmentation. Pattern Recogn. **65**, 119–135 (2017)
15. Yang, B., Zhang, X., Chen, L., Yang, H., Gao, Z.: Edge guided salient object detection. Neurocomputing **221**, 60–71 (2017)
16. Li, Z., Liu, G., Zhang, D., Xu, Y.: Robust single-object image segmentation based on salient transition region. Pattern Recogn. **52**, 317–331 (2016)
17. Yasnoff, W.A., Mui, J.K., Bacus, J.W.: Error measures for scene segmentation. Pattern Recogn. **9**(4), 217–231 (1977)
18. Fleiss, J.L., Cohen, J., Everitt, B.S.: Large sample standard errors of kappa and weighted kappa. Psychol. Bull. **72**(5), 323–327 (1969)

Machine Learning Advances and Applications (Invited Session 03)

Neuro-Fuzzy Network for PM2.5 Prediction

Yu-Chun Lin[1], Zhen-Yu Wu[1], Shie-Jue Lee[1](✉), and Chen-Sen Ouyang[2]

[1] Department of Electrical Engineering and Intelligent Electronic Commerce Research Center, National Sun Yat-Sen University, Kaohsiung, Taiwan
leesj@mail.ee.nsysu.edu.tw

[2] Department of Information Engineering, I-Shou University, Kaohsiung, Taiwan

Abstract. This paper proposes a PM2.5 prediction system based on neuro-fuzzy neural networks which can be trained through historical recorded information. Time series training data are employed to forecast the PM2.5 values in the air in the future. Because of the uncertainty of the involved impact factors, fuzzy elements are added to the forecasting system. Our prediction system is a four-layer fuzzy neural network, consisting of the input layer, fuzzy layer, inference layer, and output layer. First of all, training data are partitioned into fuzzy clusters whose membership functions are characterized by the learned means and variances. Fuzzy rules are then extracted and constructed. Next, least squares optimization and gradient descent backpropagation are applied to refine the parameters of the fuzzy rules. The output of the system, indicating the forecast PM2.5, is derived through the fuzzy inference process. Experimental results are shown to demonstrate the effectiveness of the proposed forecasting system.

Keywords: Fuzzy rule · Membership function · Gradient descent Backpropagation · Least squares

1 Introduction

Air quality is an important concern everywhere around the world. In recent years, the influence of Particulate Matter (PM) on human health has become a research and social topic. PM is a complex mixture of extremely small particles and liquid droplets that get into the air. From a medical point of view, atmospheric PM can penetrate into the respiratory system. Once inhaled, these particles can affect the heart and lungs and cause serious health effects. Long-term exposure to high concentrations of atmospheric PM can cause decreased lung function and premature death [11]. Pathological research has pointed out that PM2.5 (fine aerosols with a particle size of less than 2.5 microns in air) tend to penetrate into the gas exchange regions of the lung and may pass through the lungs to affect other organs [6,9]. For these reasons, people are paying more and more attention to the research of PM2.5.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 269–276, 2019.
https://doi.org/10.1007/978-3-030-04585-2_32

Meteorological environmental factors have an impact on the accumulation and change of PM2.5 [10]. Studies conducted show that the concentration of PM2.5 is related to meteorological factors, wind speed, surface radiation and atmospheric temperature, and is also related to SO_2, TSP, NO_2, CO, O_3, etc., of local pollution emissions [4]. Under different seasonal conditions, the impact of each meteorological factor is also different. Based on the above analysis, observing hourly or daily pollutant concentrations through established air quality monitoring systems can provide warnings to people and initiate large-scale preventive pollution management actions. Predicting the PM2.5 values in the air is still a challenge because of the complexity of the processes involved and the strong coupling between many parameters, which affects the performance of any predicting system [5].

Recently, the use of artificial neural networks (ANNs) has expanded to modeling PM2.5 contaminants, and ANN is becoming an effective and popular alternative to traditional methods [1,12]. Pérez et al. [8] predict the value of PM2.5 per hour for the next day in San Diego (Chile). They conclude that the ANN model performs better than the multiple linear regression model. Ordieres et al. [7] compare the performance of the neural network model and two classical models, continuous and multiple linear regression, in predicting the daily mean of PM2.5 concentrations. The results show that the neural method is superior to the classical model and that the use of meteorological predictors contributes to their performance.

In this paper, we propose a PM2.5 prediction system based on neuro-fuzzy neural networks which can be trained through historical recorded information. Time series training data are employed to forecast the PM2.5 values in the air in the future. Because of the uncertainty of the involved impact factors, fuzzy elements are added to the forecasting system. Our prediction system is a four-layer fuzzy neural network, consisting of the input layer, fuzzy layer, inference layer, and output layer. First of all, training data are partitioned into fuzzy clusters whose membership functions are characterized by the learned means and variances. Fuzzy rules are then extracted and constructed. Next, least squares optimization and gradient descent backpropagation are applied to refine the parameters of the fuzzy rules. The output of the system, indicating the forecast PM2.5 value, is derived through the fuzzy inference process. Experimental results are shown to demonstrate the effectiveness of the proposed forecasting system.

The rest of this paper is organized as follows. Neuro-fuzzy modeling is briefly presented in Sect. 2. Section 3 describes our proposed forecasting system in detail. Experimental results are presented in Sect. 4. Finally, a conclusion is given in Sect. 5.

2 Neuro-Fuzzy Modeling

Lee and Ouyang published a neuro-fuzzy model in [3] for function approximation. Two stages, structure identification and parameter identification, are taken. In the first stage, self-constructing clustering is applied and a set of initial fuzzy rules are generated. Each fuzzy rule is of the following form:

$$\begin{aligned}&\text{IF } x_1 \text{ IS } f_1(x_1) \text{ AND } \ldots \text{ AND } x_n \text{ IS } f_n(x_n)\\&\text{THEN } y \text{ IS } c\end{aligned} \tag{1}$$

where $f_i(x_i)$'s, $1 \leq i \leq n$, are fuzzy sets, and c is a scalar. Each fuzzy set is associated with its center and deviation, which are called input parameters. In contrast, c is called an output parameter.

In the second stage, the parameters involved in the fuzzy rules are refined by a hybrid algorithm. The output parameters are updated by the least squares method, while the input parameters are updated by the batch gradient descent method. Finally, final fuzzy rules are obtained and can be used for function approximation through fuzzy inference.

3 Proposed Approach

The main purpose of this work is to predict PM2.5 with a neuro-fuzzy prediction model. Six steps are involved: (1) Correlation calculation. The correlations between variables are calculated to sort out unrelated variables. (2) Extracting training patterns. By selecting delay coefficients, the set of training patterns is extracted from the given time series data about PM2.5. (3) Clustering training data. An iterative self-constructing clustering algorithm is applied to the training patterns, and a group of clusters, together with their centers and deviations, are obtained. (4) Getting initial fuzzy rules. A set of fuzzy rules, whose input and output parameters are initialized as the values obtained from the previous step, is constructed. (5) Optimizing parameters. In this step, the input and output parameters are optimized. Both least squares optimization and gradient descent backpropagation are applied to improve the accuracy and speed of the learning process. (6) PM2.5 prediction. The optimized fuzzy rules are used to predict PM2.5 values.

First, we calculate the correlation for each pair of variables. The correlation is calculated using the Pearson product difference correlation coefficient. The value of this coefficient is between $[-1, 1]$. The closer to 1 is, the stronger the relationship is. A threshold value is set to remove weak correlated variables.

Next, we need to decide the delay coefficient of the variables. For two variables $\mathbf{x}_1$ and $\mathbf{x}_2$, if the delay coefficient is set to q, the elements $x_{1,t}$, $x_{2,t}$, $x_{1,t-1}$, $x_{2,t-1}$, $\ldots$, $x_{1,t-q}$, $x_{2,t-q}$ are to be included and there are a total of $2q + 2$ inputs. Consider the following series of data $\mathbf{x}_1$ and $\mathbf{x}_2$:

$$x_{1,1}, x_{1,2}, \ldots, x_{1,t} \tag{2}$$

$$x_{2,1}, x_{2,2}, \ldots, x_{2,t} \tag{3}$$

Suppose we want to use $\mathbf{x}_1$ and $\mathbf{x}_2$ to predict $\mathbf{x}_1$. The training patterns can be extracted as

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_M, Y_M) \tag{4}$$

where $M = t - q$, and X_i and Y_i are as follows:

$$X_i = \{x_{1,i}, x_{2,i}, x_{1,i+1}, x_{2,i+1}, \ldots, x_{1,i+q-1}, x_{2,i+q-1}\}, \tag{5}$$

$$Y_i = x_{1,i+q} \tag{6}$$

for $1 \leq i \leq M$.

After constructing the training patterns, the training patterns are grouped into clusters with their own centers, deviations, and heights. The clustering method adopted is the iterative self-constructing clustering algorithm. When clustering is done, we have K clusters G_1, G_2, ..., G_K. Each cluster has its own center, deviation, and height.

Next, we construct a set of K initial fuzzy rules. For cluster G_j, a fuzzy rule is defined as

$$\begin{aligned}&\text{IF } x_1 \text{ IS } \mu_{1,j}(x_1) \text{ AND } \ldots \text{ AND } x_n \text{ IS } \mu_{n,j}(x_n)\\ &\text{THEN } y \text{ IS } c_j\end{aligned} \tag{7}$$

where c_j is the output parameter, and $\mu_{i,j}$ is a membership function defined as

$$\mu_{i,j}(x_i) = g(x_i; m_{i,j}; \sigma_{i,j}) = exp[-(\frac{x_i - m_{i,j}}{\sigma_{i,j}})^2] \tag{8}$$

with $m_{i,j}$ and $\sigma_{i,j}$ being the center and deviation, respectively, along the ith dimension of cluster G_j, and c_j is the height of G_j. The fuzzy operator AND is interpreted as an algebraic product, defined as

$$\prod_{i=1}^{n} g(x_i; m_{i,j}; \sigma_{i,j}). \tag{9}$$

A fuzzy network is then constructed, consisting of four layers called input layer, fuzzification layer, inference layer, and output layer, respectively. The network operates as follows.

– The input layer contains n nodes. The i-th node of the layer generates an output $o_i^{(1)}$ from the input signal p_i and sends it to the fuzzification layer, i.e.

$$o_i^{(1)} = p_i \tag{10}$$

for $1 \leq i \leq n$.

– The fuzzification layer contains K groups, each group has n nodes. The (i, j)th node produces its output $o_{i,j}^{(2)}$ by computing the corresponding Gaussian function, i.e.

$$o_{i,j}^{(2)} = \mu_{i,j}(o_i^{(1)}) = g(o_i^{(1)}; m_{i,j}; \sigma_{i,j}) \tag{11}$$

for $1 \leq i \leq n$ and $1 \leq j \leq K$.

– The inference layer contains K nodes, and the output $o_j^{(3)}$ of node j is the product of all outputs from the second layer, i.e.

$$o_j^{(3)} = \prod_{i=1}^{n} o_{i,j}^{(2)} \tag{12}$$

for $1 \leq j \leq K$.

– The output layer contains only one node, and its output $o^{(4)}$ represents the result of the centroid deblurring, i.e.

$$o^{(4)} = \frac{\sum\limits_{j=1}^{K} o_j^{(3)} c_j}{\sum\limits_{j=1}^{K} o_j^{(3)}} \tag{13}$$

The input and output parameters are to be refined later.

The next step is to optimize the fuzzy rules parameters. Assume (X_v, Y_v) is the $v$thth training pattern, X_v is the input vector, and Y_v is the desired output. The actual output can be obtained through the neural fuzzy network:

$$v.o^{(4)} = a_{v1}c_1 + a_{v2}c_2 + \cdots + a_{vK}c_K, \tag{14}$$

$$a_{vj} = \frac{v.o_j^{(3)}}{\sum\limits_{j=1}^{K} v.o_j^{(3)}}. \tag{15}$$

We want $|Y_v - v.o^{(4)}|$ to be as small as possible, so that the actual output is very close to the expected output. The objective function can be defined as follows:

$$J(C) = \|B - AC\| \tag{16}$$

where B is the matrix of expected outputs, A is the matrix of training pattern membership degrees, and C is the matrix of output parameters:

$$B = \left[Y_1 \; Y_2 \; \ldots \; Y_M\right]^T, \tag{17}$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1K} \\ a_{21} & a_{22} & \ldots & a_{2K} \\ \vdots & \vdots & \vdots & \vdots \\ a_{M1} & a_{M2} & \ldots & a_{MK} \end{bmatrix}, \tag{18}$$

$$C = \left[c_1 \; c_2 \; \ldots \; c_K\right]^T. \tag{19}$$

First, A and B are kept fixed to update C by the least squares optimization algorithm. Then C is kept fixed to update $m_{i,j}$ and $\sigma_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq K$, by the batch gradient descent method as follows. Let

$$E = \frac{1}{2N} \sum_{v=1}^{M} (Y_v - v.o^{(4)})^2 \tag{20}$$

where Y_v is the expected output and $v.o^{(4)}$ is the actual output of the vth training pattern. The updating formulae are:

$$m_{i,j}^{new} = m_{i,j}^{old} - \eta_1 (\frac{\partial E}{\partial m_{i,j}}), \tag{21}$$

$$\sigma_{i,j}^{new} = \sigma_{i,j}^{old} - \eta_2 (\frac{\partial E}{\partial \sigma_{i,j}}) \tag{22}$$

The learning process is iterated until convergence is reached.

4 Experimental Results

To evaluate the predicted results, we use Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) which are defined as

$$\text{MAE} = \frac{1}{N} \sum_{k=1}^{N} |r_k - \hat{r}_k|, \tag{23}$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_k - \hat{r}_k)^2} \tag{24}$$

where N is the number of testing patterns, and r and $\hat{r}$ are actual and predicted values, respectively.

The data set is the RNCAB dataset collected at the two stations "Kennedy" and "Parque Simon Bolivarhk" in 2015 [2]. Table 1 shows the performance of our method. Comparisons with the method of [2] are also shown in this table. Note that our method performs better than the two models, ANN-MLP and ANN-MLP-1, proposed in [2].

Table 1. Performance comparisons on PM2.5

<table>
<tr><th>Station</th><th>Model</th><th>MAE</th><th>RMSE</th></tr>
<tr><td>Parque</td><td>ANN-MLP</td><td>5.28</td><td>7.16</td></tr>
<tr><td>Simon</td><td>ANN-MLP-1</td><td>5.02</td><td>6.57</td></tr>
<tr><td>Bolivar</td><td>Our method</td><td>1.02</td><td>4.87</td></tr>
<tr><td rowspan="3">Kennedy</td><td>ANN-MLP</td><td>6.10</td><td>7.87</td></tr>
<tr><td>ANN-MLP-1</td><td>4.72</td><td>5.79</td></tr>
<tr><td>Our method</td><td>3.42</td><td>4.48</td></tr>
</table>

5 Conclusion

We have presented a PM2.5 prediction system based on neuro-fuzzy neural networks which can be trained through historical recorded information. Time series training data are employed to forecast the PM2.5 values in the air in the future. Our prediction system is a four-layer fuzzy neural network, consisting of the input layer, fuzzy layer, inference layer, and output layer. First of all, training data are partitioned into fuzzy clusters whose membership functions are characterized by the learned means and variances. Fuzzy rules are then extracted and constructed. Next, gradient descent backpropagation and least squares optimization are applied to refine the parameters of the fuzzy rules. The output of the system, indicating the forecast PM2.5 value, is derived through the fuzzy inference process. Experimental results have been shown to demonstrate the effectiveness of the proposed forecasting system.

Acknowledgment. This work was supported in part by grants MOST-104-2221-E-110-052-MY2 and MOST-105-2622-E-110-014-CC3, by the NSYSU-NUK Joint Research Project #NSYSUNUK-107-P006, by the "Intelligent Electronic Commerce Research Center" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan, and by grant B106002, STSP AI Robot Project, MOST, TAIWAN.

References

1. Abderrahim, H., Chellali, M.R., Hamou, A.: Forecasting PM10 in Algiers: efficacy of multilayer perceptron networks. Environ. Sci. Pollut. Res. **23**(2), 1634–1641 (2016)
2. Franceschi, F., Cobo, M., Figueredo, M.: Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogot, Colombia, using artificial neural networks, principal component analysis, and k-means clustering. Atmos. Pollut. Res. **9**, 912–922 (2018)
3. Lee, S.J., Ouyang, C.S.: A neuro-fuzzy system modeling with self-constructing rule generationand hybrid SVD-based learning. IEEE Trans. Fuzzy Syst. **11**(3), 341–353 (2003)
4. Zhang, L.-O., Mon-Ling Chiang, C.Y.L.: Factors affecting suspended particulate matter (PM10) - a case study of traffic air quality monitoring stations in Taiwan. J. Soil Water Conserv. **47**(1), 1235–1246 (2015)
5. Mallet, V., Sportisse, B.: Air quality modeling: from deterministic to stochastic approaches. Comput. Math. Appl. **55**(10), 2329–2337 (2008)
6. Monn, C.: Exposure assessment of air pollutants: a review on spatial heterogeneity and indoor/outdoor/personal exposure to suspended particulate matter, nitrogen dioxide and ozone. Atmos. Environ. **35**(1), 1–32 (2001)
7. Ordieres, J., Vergara, E., Capuz, R., Salazar, R.: Neural network prediction model for fine particulate matter (PM2.5) on the US-Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua). Environ. Model. Softw. **20**(5), 547–559 (2005)
8. Pérez, P., Trier, A., Reyes, J.: Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile. Atmos. Environ. **34**(8), 1189–1196 (2000)

9. Pope III, C.A., Burnett, R.T., Thun, M.J., Calle, E.E., Krewski, D., Ito, K.: Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. Jama **287**(9), 1132–1141 (2002)
10. Rodrıguez, S., Querol, X., Alastuey, A., Kallos, G., Kakaliagou, O.: Saharan dust contributions to PM10 and TSP levels in outhern and eastern Spain. Atmos. Environ. **35**(14), 2433–2447 (2001)
11. Turner, M.C., Krewski, D., Pope III, C.A., Chen, Y., Gapstur, S.M., Thun, M.J.: Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers. Am. J. Respir. Crit. Care Med. **184**(12), 1374–1381 (2011)
12. Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M.: Intercomparison of air quality data using principal component analysis, and forecasting of PM10 and PM2.5 concentrations using artificial neural networks, in Thessaloniki and Helsinki. Sci. Total Environ. **409**(7), 1266–1276 (2011)

A Comparative Study on Encoding Methods of Local Binary Patterns for Image Segmentation

Chih-Hung Wu(✉), Chih-Chin Lai, Hsien-Jen Lo, and Po-Sen Wang

Department of Electrical Engineering, National University of Kaohsiung, No. 700, Kaohsiung University Road, Nan-Tzu, Kaohsiung 811, Taiwan
{johnw,cclai}@nuk.edu.tw, my0936@gmail.com, xkp88311@gmail.com

Abstract. Clustering for image segmentation is widely used in image recognition. Euclidean distances among pixels are calculated for discriminating the belongingness to clusters. However, Euclidean distances vary in different feature representation that may affect the image clustering results. This study discusses the abovementioned problem wherein the local binary pattern (LBP) is employed as features for clustering. Four popular LBP encoding schemes are discussed and compared. Tests on medical image segmentation using the fuzzy c-means (FCM) clustering method are conducted. Experimental results show that with a proper LBP encoding scheme, a better clustering result may be obtained.

Keywords: Image segmentation · Image feature
Local binary pattern (LBP) · FCM-based clustering

1 Introduction

Image segmentation is an essential process for image recognition. Proper image segmentation can help data analysts fast identify interesting or meaningful image blocks from a huge number of, usually unstructured, image pixels. Image segmentation is to partition a given image into homogeneous and meaningful regions. Among methods of image segmentation, clustering is one of the commonly used ones [1,2]. Clustering is a kind of unsupervised learning that collects data objects with homogeneous features in a group so that the ones with heterogeneous features are separated in different groups. Centroid-based clustering methods, such as KM and FCM [3,4], discriminate the belongings of a data object to a cluster by computing the "distance" between the data object and the cluster center. For this purpose, Euclidean distances are calculated according to the positions of image pixels in the feature space and may vary in different feature representation schemes.

This work was supported in part by the STSP AI Robot Project under grant B106002 and the Ministry of Science and Technology, Taiwan, under grant MOST 107-2221-E-390-022.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 277–283, 2019.
https://doi.org/10.1007/978-3-030-04585-2_33

Among numbers of image feature representation schemes, the local binary patterns (LBP) is widely used for image clustering because of its robustness to illumination and pose variations and low computational complexity in implementation [5]. An image pixel in LBP is described by intensity discrimination among the pixels in a circular neighborhood and the central one. The weighted sum of the binary form of the comparison result is set as the LBP label of the central pixel. For image segmentation, a reasonable clustering result presents that similar LBPs encoded as neighboring integers and the corresponding Euclidean distances among such LBPs should be small, and vice versa. However, with the change of encoding schemes, the rationality of Euclidean distance may not be held and hence the clustering results may be invalid. This study discusses on the abovementioned problem and compares the effectiveness of four popular LBP encoding schemes for image clustering. The fuzzy c-means (FCM) clustering method is used. Some CT-images are tested and the results are analyzed. Experimental results show that proper LBP encoding may gain better performance of image segmentation.

2 FCM for Image Segmentation

The FCM [4] algorithm is briefed as follows. Suppose that there are N data objects to be clustered, m is a factor determining the level of cluster fuzziness. Let x_i be the i-th, $1 \le i \le N$, data object, and K the number of clusters known *in prior*. Also, let C_k be the k-th, $1 \le k \le K$, cluster and $|C_k|$ the number of data objects in the kth cluster, v_k the centroid of the k-th cluster, $\|x - y\|$ the distance between two data objects, where $x(y)$ can be a cluster centroid or a data object. Let μ_{ik} be the membership degree of x_i "belonging-to" C_k. FCM is an iterative process that assigns a data object to a cluster by considering its membership degree through minimizing the objective function as follows.

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \mu_{ik}^{m}\|x_i - v_k\|^2, m \ge 1. \tag{1}$$

FCM begins with a pre-assigned number K of clusters and randomly chooses K centroids. The objective function in Eq. (1) is minimized by updating μ_{ik} and v_k iteratively as

$$\mu_{ik} = 1 \Bigg/ \sum_{j=1}^{K} \left(\frac{\|x_i - v_k\|}{\|x_i - v_j\|} \right)^{\frac{2}{m-1}}, \tag{2}$$

$$v_k = \sum_{i=1}^{N} \mu_{ik}^{m} x_i \Bigg/ \sum_{i=1}^{N} \mu_{ik}^{m}. \tag{3}$$

The whole process stops when $\|U_{p+1} - U_p\| < \varepsilon$, where $U_p = [\mu_{ik}]$ is the matrix composed of all μ_{ik}'s in the pth iteration and ε is a threshold given by the user. If $\|U_{p+1} - U_p\| < \varepsilon$ does not hold, a round of FCM iterates with new centroids generated according to Eq. (3). For image segmentation, image pixels are data objects x_i's that are represented in specific image features, such as RGB, HSV, etc. Several rounds of FCM may be performed with various K until the number of reasonable image regions can be decided. FCM is widely used for image clustering due to its simplicity and clarity in both representation and implementation [6,7].

3 Local Binary Patterns

Given a circular neighborhood of radius R centered on pixel g_c in a gray-scale intensity image, the LBP label of g_c is defined as below.

$$LBP_{P,R}(g_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^P, \tag{4}$$

where g_p is the gray-scale intensity of the pth pixel in the circular neighborhood and $s(x)$=1, if $x \geq 0$, and $s(x)$=0, if $x < 0$. An LBP encoding example with $P = 8$ and $R = 1$ is given in Fig. 1. Because of its computational simplicity and robustness to illumination changes, LBP is widely used in image analysis. When LBP is used as the feature for image clustering, each pixel is encoded as an integer according to Eq. (4). A pixel in a gray image is represented as a 1-dimensional data object. A pixel in a color image is as a 3-dimensional data object, where each dimension denotes the LBP in R, G, or B color spaces. Clustering is performed by discriminating the similarity among pixels by their LBP codes.

The original LBP presents a simple weighting scheme for the circular neighbors around a given pixel. The neighbors can be counted from any position. The encoding scheme presented in Fig. 1 is commonly used. Nevertheless, there are several varieties that emphasis on specific texture patterns. Popular varieties of LBP schemes are presented in Fig. 2. For example, the 3×3 pixels in Fig. 1, the center pixel is encoded as 209 in LBP_0, 23 in LBP_1, 232 in LBP_2, and 46 in LBP_3.

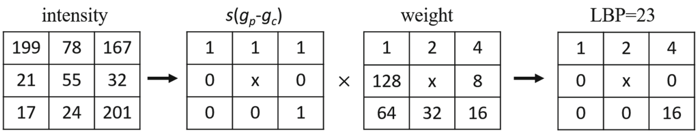


Fig. 1. An LBP encoding example

1	**128**	**64**
2	x	**32**
4	8	16

(a) LBP_0

1	2	4
128	x	8
64	**32**	16

(b) LBP_1

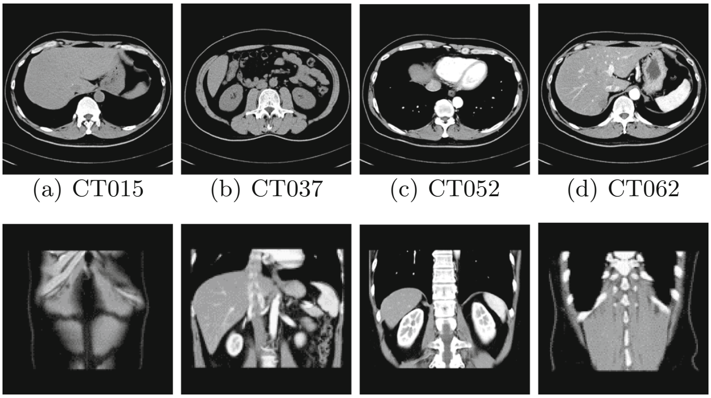
128	**64**	**32**
1	x	16
2	4	8

(c) LBP_2

2	4	8
1	x	16
128	**64**	**32**

(d) LBP_3

Fig. 2. Encoding schemes of LBP varieties



(a) CT015 (b) CT037 (c) CT052 (d) CT062

(e) CT225 (f) CT242 (g) CT251 (h) CT257

Fig. 3. Original CT images

4 Experiment

For comparing the LBP encoding schemes, the CT-images shown in Fig. 3 are used for image segmentation using FCM. These images are in 8-bit grayscale and 256×265 in size. Among these CT-images, Fig. 3(a)–(d) are cross sections and Fig. 3(d)–(f) are longitudinal sections of a human body. The FCM algorithm is implemented in C. All experiments are conducted in a PC with 8G RAM and a Core i7 CPU running on Windows 10. The K value (number of clusters) associated with FCM is set from 2 to 10. Each experiment performs for 10 times and the stable results are retained and discussed. The clustering results are evaluated by the clustering validity index (CI) which assumes that in a well-clustered image a pixel and its neighbors should belong to the same cluster. The concept of CI was introduced by [8] and extended as follows. Let L be a list of r ordered pixels, $L=\langle p_0, p_1, \ldots, p_{r-1}\rangle$. The clustering inconsistency $CI(L)$ of L is defined as

$$CI(L) = \sum_{i=0}^{r-2} \triangle\ (C(p_i), C(p_{i+1})), \tag{5}$$

where $C(p_x)$ and $C(p_y)$ are the cluster ID's of pixels p_x and p_y, respectively, and $\triangle\ (C(p_x), C(p_y))=1$ if $C(p_x) \neq C(p_y)$ and $\triangle\ (C(p_x), C(p_y))=0$ if $C(p_x) = C(p_y)$. Suppose I is an $M \times N$ image and $W_m(s,t)$ is a matrix containing $m \times m$ pixels and is centered at the position (s,t) of I, where $0 \leq s \leq M-1$, $0 \leq t \leq N-1$, and $m \leq min(M,N)$. After being clustering, the clustering inconsistency CI_2 of $W_m(s,t)$ is defined by the CI of lists of pixels extracted from rows and columns of $W_m(s,t)$.

$$CI_2(W_m(s,t)) = \sum_{i=0}^{m-1} CI(R_i) + \sum_{j=0}^{m-1} CI(C_j), \tag{6}$$

R_i and C_j are the lists of pixels extracted from the ith row and the jth column of $W_m(s,t)$, respectively, $0 \leq i, j \leq m-1$. CI_2 presents the degree of clustering inconsistency of $W_m(s,t)$. The overall clustering inconsistency $CI(I)$ can be the sum of all $CI_2(W_m(s,t))$, where there are a number of $W_m(s,t)$ selected by the users. A good clustering result may have a low CI value. In the following experiments, m=17 and all pixels in the test image I are evaluated.

(a) LBP_0 (b) LBP_1 (c) LBP_2 (d) LBP_3

Fig. 4. CT052 in various LBP encoding schemes

(a) LBP_0 (b) LBP_1 (c) LBP_2 (d) LBP_3

Fig. 5. CT225 in various LBP encoding schemes

Figures 4 and 5 present, respectively, the layouts of CT052 and CT225 in various LBP encoding schemes. By visually investigating the details of Figs. 4 and 5, the outlines of the same tissues appear differently. Table 1 lists the number of textures that are described by LBP with the high-order bits are all 1, as the gray-areas shown in Fig. 2. That is, the 5th, 6th, and 7th bits in LBP, that contribute 2^5, 2^6, and 2^7 of weights to the texture code, respectively. Textures with all bits are 0 or 1 are ignored. An interesting phenomenon is that cross-section CT-images (Fig. 3(a)–(d)) have many LBP0 and LBP_1 textures, both emphasize on the corner-like patterns. Meanwhile, longitudinal section CT-images (Fig. 3(e)–(f)) have many LBP2 and LBP_3 textures, both emphasize on the line-like patterns. These emphasized texture patterns may dominate the clustering results because they present larger Euclidean-values.

The above observations are verified by CI_2 after clustering. For each CT-image, the best (lowest) CI value appears at the clustering result where many specific LBP textures appear. The ratio of consistent LBP optimums in Tables 1 and 2, where the best CI value of a CI-image clustering appears with the most LBP patterns in the same LBP encoding scheme, is 75% (6 out of 8 cases).

Table 1. CT LBP counts

CT-Id	015	037	052	062	225	242	251	257
LBP_0	3105	2731	2452	**2564**	4366	3756	5024	3761
LBP_1	**3352**	**2845**	**2482**	2510	4126	3615	4988	3846
LBP_2	2876	2599	2235	2462	4799	**3952**	**5162**	**4106**
LBP_3	2945	2671	2265	2397	**4867**	3916	5101	3945

Table 2. Clustering inconsistency values (K: ground truth clusters)

CI-Id	015(6)	037(5)	052(6)	062(7)	225(6)	226(5)	251(7)	257(6)
LBP_0	5001	5418	**1114**	**4178**	4208	3687	4309	4987
LBP_1	**3929**	**4995**	3229	6056	3816	4199	5034	5420
LBP_2	4534	5132	2592	4551	3714	**3456**	3938	**3240**
LBP_3	4813	5136	1409	4915	**2605**	4093	**3552**	6548

5 Conclusion

This study compares the LBP encoding schemes for image clustering. Four LBP encoding schemes are tested on CT-image segmentation. The results show that a LBP encoding scheme emphasizing on a specific type of image texture dominates the feature representation of an image, and hence, varies the clustering result. Tests on CT-image segmentation present that LBP_0/LBP_1 and LBP_2/LBP_3 may be suitable for describing cross-section and longitudinal-section images, respectively. More studies should be conducted for verifying the connections between the encoding scheme and clustering results. Consistency is an essential requirement that a feature measure system should hold. When the Euclidean-distance is used as a measure for describing the homogeneity of features, it is required that pixels in similar features should be close in Euclidean-distance, and vice versa. The original LBP presents a simple weighting scheme for the circular neighbors that can be counted from any position without considering their texture variations in terms of Euclidean-distance. Optimizing the LBP encoding scheme that considers both the uniqueness of each texture and the homogeneity among textures may further provide significant feature information for image segmentation. These will be included in our future work.

References

1. Yi, L., Zhang, G., Wu, Z.: A scale-synthesis method for high spatial resolution remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens. **50**, 4062–4070 (2012)
2. Chen, J., Zheng, H., Lin, X., Wu, Y., Su, M.: A novel image segmentation method based on fast density clustering algorithm. Eng. Appl. Artif. Intell. **73**, 92–110 (2018)

3. Forgy, E.: Cluster analysis of multivariate data: efficiency versus interpretability of classification. Biometrics **21**(3), 768–769 (1965)
4. Dunn, J.C.: A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J. Cybern. **3**(3), 32–57 (1973)
5. Ojala, T., Pietikainen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recogn. **29**(1), 51–59 (1996)
6. Parida, P., Bhoi, N.: Fuzzy clustering based transition region extraction for image segmentation. Eng. Sci. Technol. Int. J. **21**(4), 547–563 (2018)
7. Abbasi, S., Tajeripour, F.: Detection of brain tumor in 3D MRI images using local binary patterns and histogram orientation gradient. Neurocomputing **219**, 526–535 (2017)
8. Awad, M., Chehdi, K., Nasri, A.: Multi-component image segmentation using a hybrid dynamic genetic algorithm and fuzzy C-means. IET Image Proc. **3**(2), 52–62 (2009)

ECG-Based Automatic Sleep Staging Using Hidden Markov Model

Shing-Tai Pan[1(✉)], Chih-Hung Wu[2], Chia-Ho Wu[1], Yung-Ran Lin[2], and Shie-Jue Lee[3]

[1] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, R.O.C.
stpan@nuk.edu.tw, m1025511@mail.nuk.edu.tw
[2] Department of Electrical Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan, R.O.C.
johnw@nuk.edu.tw, f04050714@gmail.com
[3] Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung 804, Taiwan, R.O.C.
leesj@mail.ee.nsysu.edu.tw

Abstract. In this paper, an automatic sleep stages recognition system based on the electrocardiogram (ECG) signals is developed. The reason that ECG signals are used for sleep staging is that the device for measuring ECG signals is cheap and is portable. So, the sleep staging can then be performed at home. In this study, some ECG sleep features used in other research are adopted. These features are used to train the Hidden Markov Model (HMM) model and then fed into the trained HMM for recognition. Unlike the existing research on sleep staging by HMM, in which the modeling of HMM is independent of the special properties of the sleep stage transition, the HMM in this study is adjusted to meet these properties. With this method, the accuracy of sleep staging can be improved. The experimental results show that the proposed method enhances the recognition rate compared with other existing research.

Keywords: Sleep staging · Hidden Markov Model (HMM) Electrocardiogram (ECG)

1 Introduction

In human physiology, a good deep sleep (SWS) stage can help with physical recovery; as well, a good rapid eye movement (REM) stage can improve one's learning abilities and memories. Besides, in human pathology, many studies are relating various illnesses to sleep. These include night terrors, nocturnal enuresis, sleepwalking, somniloquy, sleep apnea, and many more. Therefore, it is vital to have a more reliable sleep medical diagnosis for the patients with sleep problems [1, 2].

Sleep staging is traditionally done according to Rechtschaffen and Kales (R&K 1968) rules. In addition to manual sleep staging by doctors and experts, researches have been conducted in recent years on automatic sleep staging for the purpose of staging speed and accuracy. These researches include: application of Artificial Neural Network

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 284–291, 2019.
https://doi.org/10.1007/978-3-030-04585-2_34

(ANN) on automatic sleep staging [3], rule-based automatic sleep staging [4], and applications of HMM on automatic sleep staging [5, 6].

However, the recognition rate of sleep staging in existing literature is still low. Therefore, how to improve the recognition rate is still an important issue. Since, the HMM allows for the analysis of non-stationary multivariate time series by modeling both the state transition probabilities and the probability of observation of a state. The state estimation in HMM process depends on the preceding state. That is, in the HMM process, the result of the previous state will influence the state recognition result of the next state. This is similar to the process for sleep staging, where must consider the relationship between the previous sleep stage and the next sleep stage. Hence, the HMM is a promising model for sleep staging which possesses the properties of successive stage transition.

The selection of features is the most important step for recognition problems. A good recognition result will be obtained, if the selected features used for training and testing are good enough to discriminate between the states of time series. In this paper, some features selected from the previous research are used. The experimental results in this paper show that these features are effective in sleep staging.

2 Sleep Features from ECG Signals

The sleep features used in this paper are mainly obtained from the heart rate variability (HRV) of ECG signals. Both time-domain features and frequency-domain features calculated from HRV are commonly used in many researches for various applications. The most commonly used time-domain features and frequency-domain features are listed in Tables 1 and 2, respectively [7]. However, it is important to evaluate the features listed in Tables 1 and 2 for sleep staging problem. In fact, this work had been examined in the research [8]. Hence, in this paper, the features used for sleep staging are selected according to those in [8] and are listed in Table 3. There are two features selected from frequency domain and three features selected from time domain.

Table 1. Time-domain features calculated from HRV

Features	Definitions	Unit
SDNN	Standard Deviation of Normal to Normal	ms
SDANN	Standard Deviation of the Averages of NN intervals in all 5-min segments of the entire recording	ms
RMSSD	Square root of the mean of the sum of the squares of the successive differences between adjacent NNs	ms
NN50 count	Number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording	numbers
pNN50	NN50 count divided by the total number of all NN intervals	%

Table 2. Frequency-domain features calculated from HRV

Features	Definitions	Unit	Range of frequency
TP	Total Power (sum of the power in high frequency, low frequency and very low frequency)	ms^2	≤ 0.4 Hz
VLF	Very Low Frequency Power	ms^2	≤ 0.04 Hz
LF	Low Frequency Power	ms^2	0.04–0.15 Hz
HF	High Frequency Power	ms^2	0.15–0.4 Hz
NLF	Normalized LF = LF/(TP-VLF)	nu	
NHF	Normalized HF = HF/(TP-VLF)	nu	
RLHF	Ratio of LF and HF = LF/HF	nu	

nu: normalized unit

Table 3. The features used in this paper for sleep staging

Type of domain	No.	Features
Frequency domain	1	HF (high frequency power)
	2	LF (low frequency power)
Time domain	3	Average heart rate variability per 30 s
	4	SDNN (Standard Deviation of Normal to Normal)
	5	Average heart rate per 30 s

3 Hidden Markov Model for Sleep Staging

According to the type of the probability distributions used in HMMs, HMMS can be categorized as Continuous Hidden Markov Model (CHMM) and Discrete Hidden Markov Model (DHMM). The DHMM provides more stable recognition results and faster training with the recognition accuracy that is no less than CHMM. Therefore, the DHMM is adopted for sleep staging in this paper. Some useful features of the ECG signals are selected to training the HMM. The constructed HMM model can become a reliable computer-assisted tool for the clinical staff to increase the efficiency of sleep stage recognition in the future.

Figure 1 illustrates the training process of HMM model. First, the matrix A, B and π which describe the HMM model and will be explained in the following text, are randomized at initial step. Then, the features are quantized through the trained codebook. The quantized features are then the observation of the HMM model. The corresponding probability of the observation can be found from the present value of A, B and π. Using these probabilities, we can run the Viterbi Algorithm [9] to update the matrix A, B and π until the values in these matrix converge.

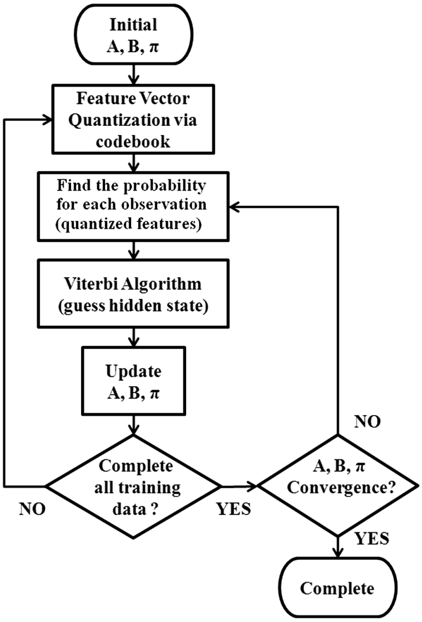


Fig. 1. The HMM model training process.

In the following, the definition and detail training method of the parameters in HMM are introduced [9]. First, the definition of parameters in HMM is introduced as follows.

λ: HMM model, $\lambda = (A, B, \pi)$

A: $A = [a_{ij}]$, a_{ij} is the probability of state x_i transferring to state x_j, $a_{ij} = P(q_t = x_j | q_{t-1} = x_i)$

B: $B = [b_j(k)]$, $b_j(k)$ is the probability of *kth* observation which is observed from the state x_i i.e., $b_j(k) = P(o_t = v_k | q_t = x_j)$

π: $\pi = [\pi_i]$, π_i is the probability of the case where the initial state is x_i, $\pi_i = P(q_1 = x_i)$

X: the state vectors of HMM, $X = (x_1, x_2, \cdots, x_N)$

V: the observation event vector of HMM, $V = (v_1, v_2, \cdots, v_M)$

O: the observation results of HMM, $O = o_1, o_2, \cdots, o_T$

Q: the resulting states of HMM, $Q = q_1, q_2, \cdots, q_T$.

To train the HMM model parameters $\lambda = (A, B, \pi)$ based on existing data, some notations are defined for convenience as follows:

E_{ij}: the event of the transition from state x_i to state x_j

$E_{i\bullet}$: the event of the transition from state x_i to other states

$E_{\bullet j}$: the event of the transition from other states to state x_j

E_{hi}: the event of state x_i appears at initial state

$n(E_{ij})$: the number of the transition from state x_i to state x_j

$n(E_{i\bullet})$:	the number of the transition from state x_i to other states
$n(E_{\bullet j})$:	the number of the transition from other states to state x_j
$n(E_{\bullet j}, o = v_k)$:	the number of enter to state x_j and observation code is v_k
$n(E_{hi})$:	the number of the event of state x_i appears at initial state.

In the process of training the matrix A, B, and π of HMM, the hidden states for each observation are estimated first through the initial A, B, and π by using Viterbi Algorithm. Then the values $n(E_{ij}), n(E_{i\bullet}), n(E_{\bullet j}), n(E_{\bullet j}, o = v_k), n(E_{hi})$ are computed according to the estimated hidden states for the whole training data. Subsequently, the elements in matrices A, B, and π are updated as follows.

$$\overline{a_{ij}} = \frac{n(E_{ij})}{n(E_{i\bullet})} \tag{1}$$

$$\overline{b_j}(k) = \frac{n(E_{\bullet j}, o = v_k)}{n(E_{\bullet j})}, \tag{2}$$

$$\overline{\pi_i} = \frac{n(E_{hi})}{n_{TD}}, \tag{3}$$

where n_{TD} is the number of training data. Using these updated A, B, and π, we run the Viterbi Algorithm again. The above steps are repeated until the matrices A, B, and π converge. The training process for a HMM is then completed. During the recognition process, the probability of the observations according to the model $\lambda(A, B, \pi)$ is calculated by the following Eq. (4) [9]:

$$\begin{aligned} P(O|\lambda) &= \sum_{Q} P(O|Q,\ \lambda)P(Q|\lambda) \\ &= \sum_{q_1 \cdots q_T} \pi_{q_1} b_{q_1}(o_1) \cdot a_{q_1 q_2} b_{q_2}(o_2) \cdots a_{q_{T-1} q_T} b_{q_T}(o_T). \end{aligned} \tag{4}$$

This equation enables us to evaluate the probability of the observations O based on the DHMM model $\lambda(A, B, \pi)$. However, the time taken to evaluate $P(O|\lambda)$ directly would be an exponential function of the observation number T. For this reason, the Forward Algorithm [9] is applied to reduce the computation time.

After introducing the HMM model, the strategy for sleep staging is then proposed. Based on the features introduced in the Sect. 2, a codebook is first trained. In the HMM recognition process, there is a HMM model should be trained according to the trained codebook and used for sleep staging. Since the transition of hidden state of HMM is similar to that of the sleep stage, the sleep stages are assigned to be the hidden stages of the HMM. Consequently, for the HMM $\lambda = (A, B, \pi)$ applied on sleep staging, the matrix A means the probability of transition from one sleep stage to another, the matrix B means the probability of the observation from the sleep stage to which the feature vector belongs, the matrix π means the probability of a sleep stage occurs in the initial stage in a sleep sample. The features of each epoch via vector quantization are viewed as the observation and are used to train HMM and estimate the sleep stage by HMM. Each observation code (feature vector) is calculated from a 30-s sample of a sleep stage.

4 Experiment Results

In the experiments, the ECG data during sleep were recorded for 7–8 h from 8 subjects of male with ages between 22 to 41 years old. The sleep efficiencies of these subjects are in the range of 80% to 90%. The sampling rate for the ECG data is 128 Hz. An epoch is read for each 30 s ECG data. There are 900 epochs read from the record of each subject. An expert is assigned for recognizing the sleep stage for every epoch based on the R&K rule [10]. In this experiment, a fifty-fifty experiment is conducted. That is, the epochs of four subjects are used for training and the epochs of the other four subjects are used for testing. There are five different combinations of these data performed in this experiment. The combinations are listed in Table 4. The numbers of epoch for each sleep stage in the five combinations are listed in Table 5.

Table 4. Subject no. in the five different combinations

Combination no.	No. of subjects used for training data and testing data	
1	Training data	No.1 No.2 No.3 No.4
	Testing data	No.5 No.6 No.7 No.8
2	Training data	No.5 No.6 No.7 No.8
	Testing data	No.1 No.2 No.3 No.4
3	Training data	No.1 No.3 No.5 No.7
	Testing data	No.2 No.4 No.6 No.8
4	Training data	No.2 No.4 No.6 No.8
	Testing data	No.1 No.3 No.5 No.7
5	Training data	No.1 No.2 No.5 No.6
	Testing data	No.3 No.4 No.7 No.8

Table 5. Numbers of epoch for each sleep stage in different combinations

Sleep stages / Epoch numbers / Combinations	wake	s1	s2	sws	rem
No.1 No.2 No.3 No.4	230	113	1802	397	1058
No.5 No.6 No.7 No.8	134	163	1747	689	867
No.1 No.3 No.5 No.7	259	193	1692	508	948
No.2 No.4 No.6 No.8	105	83	1857	578	977
No.1 No.2 No.5 No.6	130	121	1830	499	1020
No.3 No.4 No.7 No.8	234	155	1719	587	905

The recognition rates for each sleep stages defined in R&K Rule are listed in Table 6. According to this table, the sleep stages s1 get the worst recognition rate. This is due to the least samples for stage 21 between all the stages. However, this result is consistent with that in some researches. It is noted that, in order to compare the results with those in the paper [11], a new sleep stage "S_light" which is the combination of the two stages s1 and s2 and is referred to light sleep is added in the Table 6. Table 7 then compares the results in this paper to those in paper [11]. According to Table 7, it's obvious that the proposed method has got better performance for each sleep stage. The average recognition rate of the proposed paper is 67.79% which is much higher than that in paper [11] which is 57.8%.

Table 6. Recognition rates for each sleep stage

Sleep stage / Recog. rate / Combination no.	wake(%)	s1(%)	s2(%)	S_light (%) (combine s1 & s2)	sws(%)	rem(%)
1	68.657	47.239	77.790	71.1	73.295	68.973
2	80.435	45.133	88.846	86.27	60.202	63.989
3	70.476	50.605	75.067	74.02	64.879	51.279
4	69.112	52.332	76.5	74.03	51.378	66.245
5	76.923	46.452	72.949	70.76	62.692	51.05
Average	***73.1206***	***48.3522***	***78.2304***	***75.236***	***62.4892***	***60.3072***

Table 7. Comparison of recognition rates for each sleep stage and average with [11]

	wake	**S_light**	**sws**	**rem**	**Average**
Kesper et al. [11]	**63.4%**	**63.4%**	**43.3%**	**44.6%**	**57.8%**
Our method	***73.12%***	***75.24%***	***62.49%***	***60.31%***	***67.79%***

5 Conclusions

This paper applies HMM for sleep stage recognition, since the HMM models the probabilities of hidden state transition and hence is suitable to the estimation of the transition of sleep stage. Some features from ECG signals are used to train the HMM model and used for sleep stage recognition. The experimental results reveal that the proposed method can get better recognition results compared to an existing research. However, the recognition rates for the sleep stage s1 are relatively low. This is because that the recorded data for s1 is less than those of the other sleep stages. Besides, the recognition rate in this paper is not high enough for practical applications. We believe that, in the future research, to find some more distinguishable features can help the improvement of recognition rates for sleep staging.

Acknowledgement. This research work was supported by the Ministry of Science and Technology of the Republic of China under contract MOST 106-2221-E-390-020-MY2, NSYSU-NUK Joint Research Project #NSYSUNUK-107-P006 and grant B106002, STSP AI Robot Project, MOST, Taiwan.

References

1. Iber, C., Ancoli-Israel, S., Chesson, A.L., Quan, S.F.: The AASM manual for the scoring of sleep and associated events. American Academy of Sleep Medicine (2007)
2. Silber, M.H., Ancoli-Israel, S., Bonnet, M.H., et al.: The visual scoring of sleep in adults. J. Clin. Sleep Med. **3**(2), 121–131 (2007)
3. Dai, Z.F.: A sleep staging method based on single channel EEG signal. Master Paper, Department of Mechanical and Electro-Mechanical Engineering, National Sun Yat-sen University (2009)
4. Cheng, Y.S.: A rule-based automatic sleep staging method. Master Paper, Department of Computer Science and Information Engineering, National Cheng Kung University (2009)
5. Flexer, A., Dorffner, G., Sykacek, P., Rezek, I.: An automatic, continuous and probabilistic sleep stager based on a hidden markov model. Appl. Artif. Intell. **16**, 199–207 (2002)
6. Doroshenkov, L.G., Konyshev, V.A., Selishchev, S.V.: Classification of human sleep stages based on EEG processing using hidden markov models. Biomed. Eng. **41**(1), 25–28 (2007)
7. Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. Task force of the European society of cardiology and the North American society of pacing and electrophysiology. Eur. Heart J. **17**(3), 354–381 (1996)
8. Pan, S.T., Wu, Y.H.: Application of genetic algorithm on ECG-based features selection for sleep staging. Master Paper, Department of Computer Science and Information Engineering National University of Kaohsiung (2012)
9. Blunsom, P.: Hidden Markov Model. The University of Melbourne, Department of Computer Science and Software Engineering, 19 August 2004
10. Cheng, Y.S.: A rule-based automatic sleep staging method. Master Paper, Department of Computer Science and Information Engineering, National Cheng Kung University (2009)
11. Kesper, K., Canisius, S., Penzel, T., Ploch, T., Cassel, W.: ECG signal analysis for the assessment of sleep-disordered breathing and sleep pattern. Med. Biol. Eng. Comput. **50**, 1–10 (2012)

Human Gait Recognition Using GEI-Based Local Texture Descriptors

Chih-Chin Lai[1(✉)], Shing-Tai Pan[2], Tsung-Pin Wen[1], and Shie-Jue Lee[3]

[1] Department of Electrical Engineering, National University of Kaohsiung, Kaohsiung 81148, Taiwan
cclai@nuk.edu.tw

[2] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung 81148, Taiwan

[3] Department of Electrical Engineering, National Sun Yat-Sen University, Kaohsiung 80424, Taiwan

Abstract. Human gait is a useful biometric feature for human identification because it can be perceived remotely without physical contact. One critical step for human gait recognition is to accurately extract visual features. In this paper, we apply the center-symmetric local ternary pattern for feature extraction to identify the person from the gait images. The classification is performed by using a support vector machine. Experiments on the CASIA gait database (Dataset B) are given to illustrate the feasibility of the proposed approach.

Keywords: Gait recognition · Gait energy image
Center-symmetric local ternary pattern · Support vector machine

1 Introduction

Human gait is an important biometric feature, which can be used to identify humans at a distance by inspecting their walking manners. It has many real-world applications for intelligent visual-based surveillance in public areas such as banks, airports, and parking lots. A wide variety of gait-based human identification approaches have been proposed in the literature, and most of them can be divided into two broad categories: model-based and model-free approaches.

The model-based methods characterize a subject by a structural model and a motion model to analyze dynamics of gait. Cunado *et al.* [1] used Hough transform to extract the positions of arms, legs, and torso and then use articulated pendulum to match those moving body parts. By analyzing leg motion, Yam *et al.* [2] showed that how they can recognize people not only by the walking gait, but also by the running gait. Huang and Boulgouris [3] investigated the fusion of several features extracted from manually-labelled silhouettes. A novel approach for human gait recognition based on the combination of three discriminative features, i.e., the area, the gravity centre, and the orientation of each

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 292–297, 2019.
https://doi.org/10.1007/978-3-030-04585-2_35

body component, was also proposed. Tafazzoli and Safabakhsh [4] constructed movements model based on anatomical proportions; then, Fourier transform was used to analyze human walking. In contrast to the model-based methods, the model-free methods consider the motion of human body holistically, and features are computed without considering the underlying structure. There exists significant work in the field of model-free approach. Han and Bhanu [5] proposed a new spatial-temporal gait representation, called the gait energy image (GEI), to characterize human walking properties for individual recognition. Bashir *et al.* [6] proposed a new gait representation termed gait entropy image (GEnI) which can be used to capture mostly motion information. Choudhury and Tjahjadi [7] introduced a gait recognition method that combines the spatio-temporal shape features of a subject's silhouettes with the subject's dynamic motion characteristics over a gait period. The proposed method considers both model-free and model-based approaches to achieve robustness against the maximum number of challenging factors of gait recognition. Lee *et al.* [8] proposed a time-sliced averaged motion history image alongside the histograms of oriented gradients to generate gait signatures in the gait recognition problem. The time-sliced multi-composite image representations can encode richer gait temporal information than single composite image representation. El-Alfy *et al.* [9] developed a new approach for gait recognition that combines the distance transform with curvatures of local contours. Their method encodes both body shapes and boundary curvatures into a novel feature descriptor that is more robust than existing gait representations.

In this paper, we propose a model-free approach which uses the center-symmetric local ternary pattern (CS-LTP) [10] to extract the local texture features of a gait energy image. The recognition performance is evaluated using the CASIA database with a support vector machine (SVM) classifier. Experimental results are provided to illustrate the feasibility of the proposed approach.

2 The Proposed Approach

The feature extraction plays a key role in the gait recognition system. In this paper, the most discriminative feature, CS-LTP, was applied to a gait energy image to extract the fine information from the image locally. Figure 1 illustrates the proposed feature extraction framework.

2.1 Gait Representation

There have been a number of gait representation techniques proposed in the literature. In this paper, we extract as much information from the GEI. Given a human walking video sequence, the human silhouettes are extracted and the size normalization and horizontal alignment are applied to each extracted silhouette image. Therefore, the gray-level GEI [5] is defined as follows:

$$G(x, y) = \frac{1}{N} B(x, y, t), \tag{1}$$

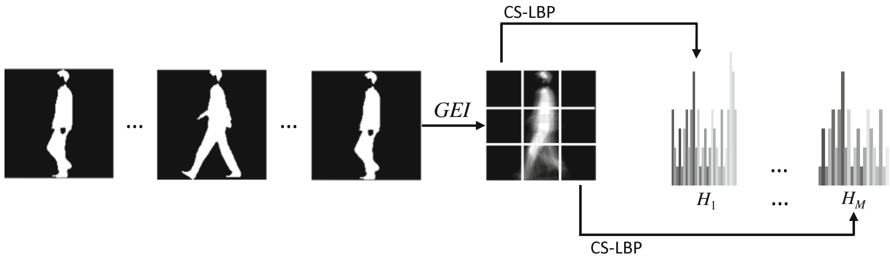


Fig. 1. The proposed GEI-based feature extraction framework

where N is the number of frames in the gait sequence, x and y are the image coordinates, and $B(x, y, t)$ is the gait silhouette image at time t in the gait sequence.

2.2 Center-Symmetric Local Ternary Pattern

The CS-LBP operator was originally designed for interest region description [11]. In contrast to LBP, CS-LBP not only has the similar desirable properties, but also captures the gradient information. In particular, CS-LBP has higher stability in flat image regions as well as significantly reduces feature dimensionality while preserving distinctiveness. The CS-LBP operator is defined as

$$CS - LBP_{N,R} = \sum_{i=0}^{N/2-1} S(n_i - n_{i+N/2})2^i, \tag{2}$$

$$S(z) = \begin{cases} 1, & \text{if } z \geq T \\ 0, & \text{otherwise} \end{cases}, \tag{3}$$

where n_i and $n_{i+N/2}$ correspond to the gray values of center-symmetric pairs of pixels of N equally spaced pixels on a circle radius R, and T is a threshold value.

The local ternary pattern (LTP) [12] was proposed to tackle the image noise in uniform regions. Instead of binary code used in the LBP, the pixel difference is encoded as a three-valued code based on a threshold. According to the similar concept of CS-LBP, Gupta *et al.* [10] proposed the CS-LTP operator, which is defined as

$$CS - LTP_{N,R} = \sum_{i=0}^{N/2-1} S(n_i - n_{i+N/2})3^i, \tag{4}$$

$$S(x) = \begin{cases} 2, x \geq T_1 \\ 1, -T_1 < x < T_1 \\ 0, x \leq -T_1 \end{cases}, \tag{5}$$

where n_i and $n_{i+N/2}$ correspond to the gray values of center-symmetric pairs of pixels of N equally spaced pixels on a circle radius R, and T_1 is a threshold value.

We divide a GEI into M non-overlapping regions $\{R_1, R_2, \ldots, R_M\}$ and then apply the CS-LTP operator on every pixel in every region. The histogram of jth region is defined as:

$$H_j = \{h_{0,j}, h_{1,j}, \ldots, h_{80,j}\}. \tag{6}$$

The histogram for every region is calculated, and then all H_j together are concatenated to obtain the final CS-LTP feature:

$$V = \{H_1, H_2, \ldots, H_M\}. \tag{7}$$

After feature extraction, the next task is to classify the different input patterns into distinct defined classes with a proper classifier. A SVM is a very popular technique for data classification in the machine learning community. The concepts of behind it are Statistical Learning Theory and Structural Minimization Principle [13]. SVM has been shown to be very effective because it has the ability to find the optimal separating hyperplane that gives the maximum margin between the positive and negative samples.

Given a training set of labeled samples $\{(x_i, y_i), i = 1, \ldots, h\}$, where $x_i \in \mathcal{R}^n$ and $y_i \in \{+1, -1\}$, a new test data x is classified by:

$$f(x) = \text{sign}\left(\sum_{i=1}^{h} \varphi_i y_i K(x_i, x) + c\right), \tag{8}$$

where φ_i are Lagrange multipliers of the dual optimization problem, c is a bias or threshold parameter, and $K(\cdot, \cdot)$ is a kernel function. In our work, we used the SVM implementation in the public available library LIBSVM [14] in all experiments.

3 Experimental Results

In the experiments, CASIA gait database (Dataset B) [15] is used to evaluate the performance of the proposed method. Dataset B consists of the data from 124 subjects under 11 viewing angles from 0° to 180°. There are six video sequences for each subject, in our experiments we have used only side view of the normal gait.

3.1 The Impact of Threshold Value and Block Number

In our approach, the threshold value of T_1 and the block number are two important factors that influence the recognition quality. In this experiment, we want to show that the appropriate experiment setups are necessary to obtain better recognition rate. The results are presented in Table 1. It can be seen that if the threshold value is fixed, the recognition rate is improved as the block number are increased. Theoretically, more blocks considered should provide more texture information; however, as the number of blocks increased, the computation effort should be increased. From our observation, in order to satisfy both recognition accuracy and computation effort requirements, $T_1 = 3$ and 10×10 blocks are the appropriate parameter values to provide better recognition performance.

Table 1. Recognition rates (%) with different threshold value and block numbers

Threshold value T_1	Block numbers							
	2×2	4×4	6×6	8×8	10×10	12×12	14×14	16×16
1	78.6	79.0	88.7	92.7	95.2	93.1	94.4	93.5
3	80.6	87.5	90.8	92.7	95.6	94.4	94.8	94.0
5	75.6	84.7	88.3	92.7	94.8	94.4	94.0	93.1
7	75.4	80.6	86.3	91.9	94.8	93.1	93.5	94.2
9	72.6	81.0	88.3	92.7	94.4	94.0	94.0	94.8
11	75.8	81.0	85.1	92.3	94.4	93.5	94.0	95.2
13	75.6	80.6	84.3	92.3	94.8	93.1	93.5	94.0

3.2 Comparison with Other Methods

In order to show the effectiveness of the proposed approach, we compare the proposed approach with those in [16–18]. Note that the results of different methods may not be directly comparable because of differences in the number of selected images, experimental setups and pre-processing procedures, but they can still indicate the discriminative performance of every approach. The experimental results are shown in Table 2. As expected, our approach can obtain satisfactory performance in comparison with other existing gait recognition methods.

Table 2. Performance comparison of different gait recognition methods

Method	Recognition rate (%)
Dupuis *et al.* [16]	77.96
Mohan Kumar and Nagendraswamy [17]	88.99
Lishani *et al.* [18]	92.06
Ours	95.60

4 Conclusions

In this paper, an approach for human identification was presented. The center-symmetric local ternary patterns are used to extract the GEI-based image features. We also use a support vector machine for gait recognition purpose. The experiments conducted on CASIA gait database have shown the good performance of the proposed approach. Further work of considering other multi-scale texture features into our approach is in progress.

Acknowledgments. This work was supported in part by grants MOST 106-2221-E-390-023 and 107-2221-E-390-019-MY2, by the NSYSU-NUK Joint Research Project #NSYSUNUK-107-P006, and by grant B106002, STSP AI Robot Project, MOST, Taiwan.

References

1. Cunado, D., Nixon, M.S., Carter, J.N.: Automatic extraction and description of human gait models for recognition purposes. Comput. Vis. Image Underst. **90**, 1–41 (2003)
2. Yam, C., Nixon, M.S., Carter, J.N.: Automated person recognition by walking andrunning via model-based approaches. Pattern Recogn. **37**, 1057–1072 (2004)
3. Huang, X., Boulgouris, N.V.: Model-based human gait recognition using fusion of features. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1469–1472. IEEE Press (2009)
4. Tafazzoli, F., Safabakhsh, R.: Model-based human gait recognition using leg and arm movements. Eng. Appl. Artif. Intell. **23**, 1237–1246 (2010)
5. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Trans. Pattern Anal. Mach. Intell. **28**, 316–322 (2006)
6. Bashir, K., Xiang, T., Gong, S.: Gait recognition using gait entropy image. In: 3rd International Conference on Imaging for Crime Detection and Prevention, pp. 1–6. IET Press (2009)
7. Choudhury, S.D., Tjahjadi, T.: Gait recognition based on shape and motion analysis of silhouette contours. Comput. Vis. Image Underst. **117**, 1770–1785 (2013)
8. Lee, C.P., Tan, A.W.C., Tan, S.C.: Time-sliced averaged motion history image for gait recognition. J. Vis. Commun. Image Represent. **25**, 822–826 (2014)
9. El-Alfy, H., Mitsugami, I., Yagi, Y.: Gait recognition based on normal distance maps. IEEE Trans. Cybern. **48**, 1526–1539 (2018)
10. Gupta, R., Patil, H., Mittal, A.: Robust order-based methods for feature description. In: 23rd IEEE Conference on Computer Vision and Pattern Recognition, pp. 334–341. IEEE Press, San Francisco (2010)
11. Heikkila, M., Pietikäinen, M., Schmid, C.: Description of interest regions with local binary patterns. Pattern Recogn. **42**, 425–436 (2009)
12. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. Image Process. **19**, 1635–1650 (2010)
13. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
14. Chang, C.-C., Lin, C.-J.: LIBSVM: a libary for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
15. Zheng, S., Zhang, J., Huang, K., He, R., Tan, T.: Robust view transformation model for gait recognition. In: 2011 18th IEEE International Conference on Image Processing, pp. 2073–2076. IEEE Press, Brussels (2011)
16. Dupuis, Y., Savatier, X., Vasseur, P.: Feature subset selection applied to model-free gait recognition. Image Vis. Comput. **31**, 580–591 (2013)
17. Mohan Kumar, H.P., Nagendraswamy, H.S.: Symbolic representation and recognition of gait: an approach based on LBP of split gait energy images. Sig. Image Process.: Int. J. **5**, 15–28 (2014)
18. Lishani, A.O., Boubchir, L., Khalifa, E., Bouridane, A.: Human gait recognition using GEI-based local multi-scale feature descriptors. Multimedia Tools Appl. (2018)

Feature Extraction of EEG Signals for Epileptic Seizure Prediction

Chen-Sen Ouyang[1](✉), Bo-Jhong Chen[1], Zong-En Cai[1], Lung-Chang Lin[2], Rong-Ching Wu[3], Ching-Tai Chiang[4], and Rei-Cheng Yang[1]

[1] Department of Information Engineering, I-Shou University, Kaohsiung 84001, Taiwan
ouyangcs@isu.edu.tw

[2] Department of Pediatrics, School of Medicine, College of Medicine, Departments of Pediatrics, Kaohsiung Medical University Hospital, Kaohsiung Medical University, Kaohsiung 80708, Taiwan

[3] Department of Electrical Engineering, I-Shou University, Kaohsiung 84001, Taiwan

[4] Department of Computer and Communication, National Pingtung University, Pingtung 90003, Taiwan

Abstract. Feature extraction of electroencephalography (EEG) signals is crucial for epileptic seizure prediction. In this study, nine scalp EEG temporal feature descriptors based on the positive zero-crossing interval length series and statistical analysis are proposed for seizure prediction. Experimental results have shown that all feature descriptors present statistical significance for discriminating pre-ictal and inter-ictal EEG epochs in a majority of subjects. Moreover, the combination of our approach and support vector machine achieves the best performance with the sensitivity 92.75%, specificity 69.77%, Kappa 16.30%, and accuracy 86.50%.

Keywords: Electroencephalography (EEG) · Feature extraction
Seizure prediction · Zero-crossing interval · Statistical analysis
Support vector machine (SVM)

1 Introduction

Epilepsy is one of most common and devastating neurologic diseases, affecting more than 50 million individuals worldwide. Up to now, many types of treatments have been developed for epilepsy, however, about 30–40% of patients cannot be controlled well by anti-epileptic drugs or surgery. The life quality of epileptics is seriously affected by the occurrence of abrupt and unpredictable seizures, which cause morbidity and mortality. Therefore, seizure prediction has become an important and urgent issue for taking preventive actions in advance to decrease the morbidity and mortality, and greatly improve the life quality of epileptics.

Physicians who care for patients with epilepsy have recognized that some patients are aware of periods when seizures will happen. Rajna et al. [6] report that clinical prodromes or auras occur in more than 50% of patients. However, children patients are less definite to describe an aura. There are also emerging evidences showing that seizures

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 298–303, 2019.
https://doi.org/10.1007/978-3-030-04585-2_36

are predictable by physiologic signals [3]. Among these, electroencephalogram (EEG) is the best and most commonly used signal in predicting seizures. For seizure prediction, some approaches [2, 8] employ time-domain analysis of EEG signals, while others employ frequency-domain analysis of EEG signals and show changes in frequencies before seizure onset [4, 7]. A majority of seizure prediction approaches are based on intracranial EEG recordings. Although intracranial EEGs can minimize environmental noise and artifacts, intracranial devices are more subject to infection and other adverse effects than scalp EEGs [1]. In this study, an efficient, quantitative, and individualized approach is proposed for pre-ictal/inter-ictal stage identification with the time-domain analysis of scalp EEG signals.

2 Our Approach

In this study, nine scalp EEG temporal feature descriptors based on the positive zero-crossing interval length series and statistical analysis are proposed for seizure prediction. For each subject with epilepsy, EEG epochs are sampled from pre-ictal and inter-ictal periods of EEG recordings by a sliding window approach. Then, a positive zero-crossing interval length series is generated from each channel signal of each EEG epoch and characterized by nine temporal feature descriptors. After that, the two-sample Kolmogorov–Smirnov test [5] is applied to evaluate the statistical significance of each descriptor. Finally, five classifiers, including Bayes, random forests, linear discrimination analysis, support vector machines, and k nearest neighbors, are trained to demonstrate the performance of seizure prediction of our approach and the comparison with the other approach is also made. The details about EEG epoch sampling and EEG feature extraction in our approach are described as follows.

2.1 EEG Epoch Sampling

For each EEG recording of each subject with epilepsy, the pre-ictal and inter-ictal periods are labeled according to the seizure event records. A pre-ictal period is defined as the 5-min interval priori to a seizure onset and has no overlap with the ictal period of previous seizure. An inter-ictal period is defined as the interval which is up to 15 min far from the ictal period of previous seizure and at least 60 min before the ictal period of next seizure [9]. Note that an ictal period is defined as the interval of a seizure event. Then, the sliding window approach is applied to each pre-ictal or inter-ictal period of EEG recording to sample a set of EEG epochs with the corresponding labels (pre-ictal or inter-ictal). Therefore, two sets, pre-ictal and inter-ictal, of EEG epochs are obtained from EEG recordings of each subject.

2.2 EEG Feature Extraction

Epilepsy is a common and devastating neurologic disease, affecting more than 50 million people worldwide. 1 Although many treatments have been developed in epilepsy recently, about 30–40% of patients cannot be controlled well by anti-epileptic drugs or by surgery. 2 Sudden and unpredictable seizures significantly affect the quality of life of these patients, and cause morbidity and mortality. 3, 4 Thus, a method that can predict seizures ahead to allow preventive action could reduce the morbidity and mortality in these patients, and greatly improve the quality of life. The most important part of seizure prediction is classification the pre-ictal and inter-ictal stages.

To characterize each channel signal in an EEG epoch with representative features, a series of positive zero-crossing interval lengths are generated and the corresponding features are extracted by nine temporal feature descriptors. Firstly, a detrending approach is applied for removing the corresponding mean and linear trend of the considered channel signal. Then, positive zero-crossing points on the time domain of the signal are identified for calculating the corresponding positive zero-crossing interval lengths. Assume $N + 1$ positive zero-crossing points are identified and the corresponding time series is $t_1, t_2, \ldots, t_{N+1}$. The positive zero-crossing interval length series is calculated as $t_2 - t_1, t_3 - t_2, \ldots, t_{N+1} - t_N$ and can also be represented as $\Delta T = \{\Delta t_n | \Delta t_n = t_{n+1} - t_n, n = 1, 2, \ldots, N\}$. After that, temporal features of the zero-crossing interval length series are extracted by nine descriptors, including *SD*, *MSD*, *RMSSD*, *pSD*50, *BLH*, *SDSD*, *IR*, *MAD*, *TI*. Definitions of all descriptors are described as follows.

1. *SD* is defined as the standard deviation of all positive zero-crossing interval lengths and calculated by

$$SD(\Delta T) = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N} \left(\Delta t_n - \overline{\Delta t}\right)^2} \tag{1}$$

   where $\overline{\Delta t}$ denotes the mean of all positive zero-crossing interval lengths.
2. *MSD* is defined as the mean of standard deviations of positive zero-crossing interval lengths in a time period and calculated by

$$MSD(\Delta T) = \frac{1}{M} \sum_{i=1}^{M} SD\left(\Delta T_i\right) \tag{2}$$

   where $SD\left(\Delta T_i\right)$ denotes the standard deviation of positive zero-crossing interval lengths during the ith time period.
3. *RMSSD* is defined as the root mean square of successive interval length differences and calculated by

$$RMSSD(\Delta T) = \sqrt{\frac{1}{N-1} \sum_{n=1}^{N-1} \left(\Delta^2 t_n\right)^2} \tag{3}$$

   where $\Delta^2 t_n = \Delta t_{n+1} - \Delta t_n, n = 1, 2, \ldots, N - 1$.
4. *pSD*50 is defined as the percentage of successive interval lengths that differ by more than 50 ms and calculated by

$$pSD50(\Delta T) = \frac{\left|\left\{\Delta^2 t_n | \Delta^2 t_n > 50\,\mathrm{ms},\ n = 1, 2, \ldots, N-1\right\}\right|}{N-1} \tag{4}$$

where $|\cdot|$ is the operator for calculating the cardinality of a set.

5. *BLH* is defined as the baseline width of the histogram of all interval lengths.
6. *SDSD* is defined as the standard deviation of successive interval length differences and calculated by

$$SDSD(\Delta T) = \sqrt{\frac{1}{N-2}\sum\nolimits_{n=1}^{N-1}\left(\Delta^2 t_n - \overline{\Delta^2 t}\right)^2} \tag{5}$$

where $\overline{\Delta^2 t}$ denotes the average of successive interval length differences.

7. *IR* is defined as the interquartile range of all interval lengths.
8. *MAD* is defined as the median of successive interval length absolute differences and calculated by

$$MAD(\Delta T) = \mathrm{median}\{|\Delta^2 t_n||n = 1, 2, \ldots, N-1\} \tag{6}$$

9. *TI* is defined as the integral of the density of the interval length histogram divided by its height and is the abbreviation of triangular index.

3 Experimental Results

In experiments, the well-known scalp EEG database, namely CHB-MIT, is adopt as the benchmark. The CHB-MIT database is composed of EEG recordings of 22 pediatric subjects with intractable seizures. A majority of EEG recording are made with 23 channels according to the international 10–20 system. All signals were sampled at 256 samples per second with 16-bit resolution. In this study, EEG recordings of 11 subjects with multiple seizures are selected from the CHB-MIT database for generating training samples and test samples. For a subject with k seizures, the pre-ictal periods and inter-ictal periods corresponding to the pervious $\lceil k \rceil$ seizures are selected for generating training samples, while the pre-ictal periods and inter-ictal periods corresponding to the remainder seizures are selected for generating test samples. Moreover, the window size is set as 30 s experimentally. To demonstrate the performance of our approach, we compare results of our approach with those of Zandi's approach [9]. In Zandi's approach, the statistical histogram is adopted as the feature descriptor of positive zero-crossing interval lengths and the window size is set as 15 s.

The p values of two-sample Kolmogorov–Smirnov test on nine proposed features of training samples of all subjects are listed in the Table 1. Note that the p values which is smaller than or equal to 0.05 are marked with bold format. Obviously, *SD*, *MSD*, *pSD*50, *BLH*, each of which presents statistical significance in each of 11 subjects are the best feature descriptors for discriminating pre-ictal and inter-ictal EEG epochs.

Table 1. The p values of two-sample Kolmogorov–Smirnov test on nine proposed features of training samples of all subjects.

ID	*SD*	*MSD*	*RMSSD*	*pSD50*	*BLH*	*SDSD*	*IR*	*MAD*	*TI*
CHB01	**0.0005**	**0.0012**	**0.0000**	**0.0005**	**0.0005**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
CHB02	**0.0130**	**0.0311**	0.0676	**0.0111**	**0.0111**	0.0555	0.0593	**0.0421**	**0.0421**
CHB03	**0.0068**	**0.0027**	**0.0000**	**0.0014**	**0.0014**	**0.0250**	**0.0000**	**0.0001**	**0.0001**
CHB04	**0.0011**	**0.0001**	**0.0014**	**0.0001**	**0.0001**	**0.0012**	**0.0009**	0.0745	0.0745
CHB05	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
CHB06	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	0.1378	0.1378
CHB07	**0.0420**	**0.0421**	**0.0000**	**0.0000**	**0.0000**	**0.0111**	**0.0000**	**0.0000**	**0.0000**
CHB08	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**
CHB09	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0319**	**0.0319**
CHB10	**0.0002**	**0.0003**	**0.0001**	**0.0002**	**0.0002**	**0.0000**	**0.0000**	0.1869	0.1869
CHB11	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0034**	**0.0034**

The classification sensitivity, specificity, Kappa, and accuracy produced by Zandi's approach and our approach for test samples are listed in Table 2. For each case of classifier and performance index, the best value of performance index is marked with bold format. Based on the overall performance on the four indices, the combination of our approach and SVM achieves the best performance with the sensitivity 92.75%, specificity 69.77%, Kappa 16.30%, and accuracy 86.50%. Especially, specificity of our approach is significantly larger than that of Zandi's approach in each case.

Table 2. The classification sensitivity, specificity, Kappa, and accuracy produced by Zandi's approach and our approach for test samples.

Classifier	Approach	Sensitivity	Specificity	Kappa	Accuracy
Bayes	Zandi et al.	**94.35**	26.70	**15.66**	**77.44**
	Ours	93.47	**32.58**	14.55	73.49
Random forests	Zandi et al.	**92.51**	26.22	10.81	**85.26**
	Ours	90.07	**31.61**	**12.05**	77.97
Linear discrimination analysis	Zandi et al.	**92.45**	26.66	**9.12**	**91.03**
	Ours	91.29	**30.30**	1.86	87.08
Support vector machines	Zandi et al.	91.92	11.41	1.01	**89.67**
	Ours	**92.75**	**69.77**	**16.30**	86.50
k nearest neighbors	Zandi et al.	91.56	23.03	6.26	**89.88**
	Ours	**92.16**	**63.91**	**14.06**	83.70

4 Conclusions

In this study, nine scalp EEG temporal feature descriptors based on the positive zero-crossing interval length series and statistical analysis are proposed for seizure prediction. Experimental results have shown that all feature descriptors present statistical significance for discriminating pre-ictal and inter-ictal EEG epochs in a majority of subjects. Moreover, the combination of our approach and SVM achieves the best performance with the sensitivity 92.75%, specificity 69.77%, Kappa 16.30%, and accuracy 86.50%. However, the number of enrolled subjects in this study is limited. More subjects will be enrolled additionally in the future works.

Acknowledgement. This work was supported by grants MOST 106-2221-E-214-048, MOST 107-2221-E-214-028, and B106002, STSP AI Robot Project, the Ministry of Science and Technology, Taiwan.

References

1. Cook, M.J., O'Brien, T.J., Berkovic, S.F., et al.: Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: a first-in-man study. Lancet Neurol. **12**(6), 563–571 (2013)
2. Lange, H.H., Lieb, J.P., Engel Jr., J., et al.: Temporo-spatial patterns of pre-ictal spike activity in human temporal lobe epilepsy. Electroencephalogr. Clin. Neurophysiol. **56**(6), 543–555 (1983)
3. Litt, B., Echauz, J.: Prediction of epileptic seizures. Lancet Neurol. **1**(1), 22–30 (2002)
4. Litt, B., Esteller, R., Echauz, J., et al.: Epileptic seizures may begin hours in advance of clinical onset: a report of five patients. Neuron **30**(1), 51–64 (2001)
5. Massey, F.J.: The Kolmogorov-Smirnov test for goodness of fit. J. Am. Stat. Assoc. **46**(253), 68–78 (1951)
6. Rajna, P., Clemens, B., Csibri, E., et al.: Hungarian multicentre epidemiologic study of the warning and initial symptoms (prodrome, aura) of epileptic seizures. Seizure **6**(5), 361–368 (1997)
7. Schiff, S.J., Colella, D., Jacyna, G.M., et al.: Brain chirps: spectrographic signatures of epileptic seizures. Clin. Neurophysiol. **111**(6), 953–958 (2000)
8. Sherwin, I.: Interictal-ictal transition in the feline penicillin epileptogenic focus. Electroencephalogr. Clin. Neurophysiol. **45**(4), 525–534 (1978)
9. Zandi, A.S., Tafreshi, R., Javidan, M., Dumont, G.A.: Predicting epileptic seizures in scalp EEG based on a variational bayesian gaussian mixture model of zero-crossing intervals. IEEE Trans. Biomed. Eng. **60**(5), 1401–1413 (2013)

Using Selective Search and CNN for Counting Motorcycles in Images

Tzung-Pei Hong[1,2](✉), Yu-Chiao Yang[1], and Ja-Hwung Su[3]

[1] Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, Taiwan
tphong@nuk.edu.tw
[2] Department of Computer Science and Engineering, National Sun Yat-sen University, Kaohsiung, Taiwan
[3] Department of Information Management, Cheng Shiu University, Kaohsiung, Taiwan

Abstract. In this paper, we propose a method for counting motorcycles from images based on the selective search and deep learning. In the proposed approach, the objects in an image are segmented by the selective searching algorithm and then recognized individually by the Convolutional Neural Network (CNN). Finally, the objects recognized as motorcycles are counted. The experimental results show the proposed method is effective in counting the motorcycles.

Keywords: Deep learning · Convolutional neural network · Selective searching Motorcycle counting

1 Introduction

Currently, there have been many researches in cars detection and they got good performances. Compared with car detection, motorcycles are more difficult to detect because there are different numbers of people on motorcycles, and their shapes are not as regular as cars. To know the number of the motorcycles on a road, manually counting is straight forward. However, it requires a high priced cost. To deal with this issue, an effective and efficient approach to automatically detect the number of motorcycles on a road is necessary.

In the past, a number of traditional researches were made on this issue by SIFT [3] or HOG [8]. Yet, the good model is not easy to approximate because of visual complexities of motorcycles. In this paper, we propose an effective method for recognizing and counting motorcycles by integrating deep learning and selective search. Firstly, the single motorcycle images are used to train the Convolutional Neural Network (CNN) for building a classification model for a motorcycle. Secondly, the object regions are extracted from the motorcycle image by the selective search algorithm [10]. Each region is then judged for whether there exists a motorcycle or not by the trained CNN model. Next, the regions with motorcycles are checked for overlapping and redundancy. Finally, the remaining regions are recognized by the count of the motorcycles in an image. The experimental evaluations reveal that the high accuracy of counting motorcycles can be achieved by the proposed idea.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 304–308, 2019.
https://doi.org/10.1007/978-3-030-04585-2_37

2 Related Work

The convolutional neural network (CNN) is a kind of deep neural network. A CNN consists of convolution layers, pooling layers and fully-connected layers [6] as shown in Fig. 1. The CNN model has been employed in many applications.

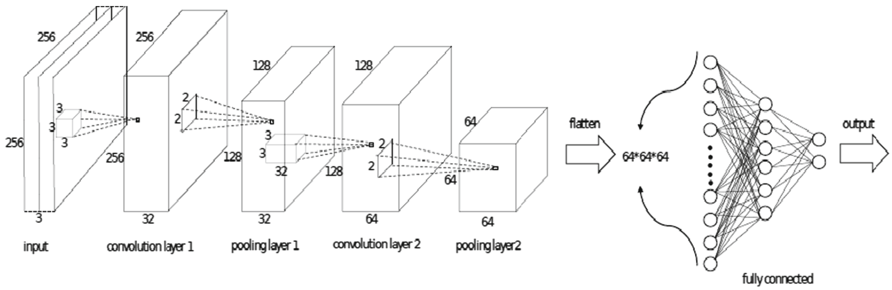


Fig. 1. The CNN architecture

In recent years, vehicle recognition has become more and more important because many related applications have been developed, such as intelligent transportation system, traffic management system, intelligent vehicle surveillance system and so on. Matthews et al. [7] used the principal component analysis (PCA) for extracting features first and thereby the multilayer perceptron (MLP) was employed for detecting vehicles by a binary classification in a sub-region of the input image. Kato et al. [5] proposed the multi-clustered modified quadratic discriminant function (MC-MQDF) for identifying vehicle position in an image. In these years, deep learning has been commonly used in pattern recognition since it has high classification accuracy [12]. For example, Chung and Sohn used the deep CNN to count the number of vehicles on a road [4].

Compared with vehicle recognition, motorcycles are more difficult to recognize. The shape of various motorcycles and the number of passengers on the motorcycle make it look different. This situation makes machine learning difficult to identify the motorcycles in an image. To identify motorcycle positions in an image, Wen et al. [11] utilized the Support Vector Machine (SVM) ensemble approach to deal with the imbalanced datasets. Mukhtar and Tang [8] used the histogram of oriented gradients (HOG) to classify the images using the SVM. Silva et al. [9] also used the SVM by Local Binary Pattern. There are still many effective approaches in development.

3 Proposed Method

In this paper, we use the deep learning model to recognize and count motorcycles in an image. The flowchart of the proposed approach is given in Fig. 2. Some important steps are stated below.

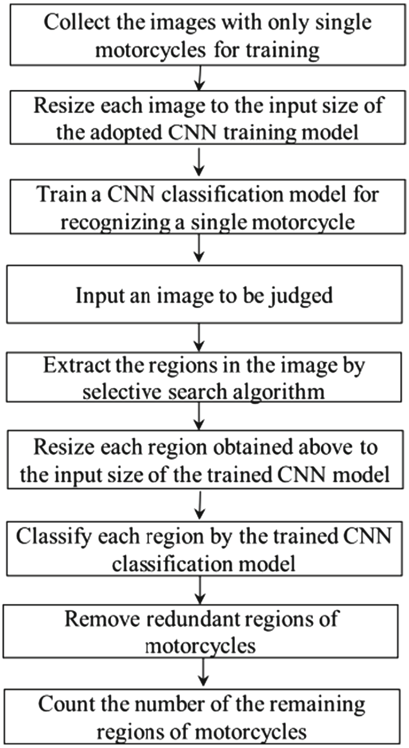


Fig. 2. The flowchart of the proposed approach

- Train a CNN model for recognizing a single motorcycle: The images of single motorcycles collected from California Institute of Technology [1] are used to train the model. In overall, the adopted CNN model includes 6 convolution layers, 3 max-pooling layers and a fully-connected layer with 3 hidden layers. Each two convolution layers are accompanied with a pooling layer. There are two final output nodes at the end of the CNN. One represents the class of motorcycles and the other denotes the class of non-motorcycles.
- Extract the regions in an image: In this step, the selective search algorithm is performed to generate the regions from an image. Each region represents an object in the image.
- Resize each region to the input size of the trained CNN model: Because the region size to be judged may not be the same as that of the input of the trained CNN model, it must be resized to fit the model first.
- Classify each region with the trained CNN: After the regions in an image are resized, they are then judged by the trained CNN model one by one.
- Remove redundant regions of motorcycles: The intersection over union (*IoU*) is adopted to measure the overlap degree of two regions. The *IoU* value [13] is denoted in Eq. (1), where *A* and *B* denote two different regions:

$$IoU(A, B) = \frac{A \cap B}{A \cup B} \tag{1}$$

The IoU value of A and B means the overlap ratio of A and B. If the measure value is larger than the overlapping threshold t, then the two regions are identified as redundant and the one with the smaller CNN output score is removed.

- Count the number of the remaining regions: the regions are counted as the number of the motorcycles.

4 Experiments

In the experiments, there are two datasets for experiments. The first dataset came from California Institute of Technology [1]. This dataset contains 826 images of a single motorcycle. The second dataset was generated by a six-minute video from the surveillance video in Ho Chi Minh City [2]. To evaluate the proposed method, a measure for error rate in motorcycle counting is proposed as shown in Eq. (2):

$$counting\ error\ rate = \begin{cases} \dfrac{|A - P|}{A}, & if\ A \neq 0 \\ P, & if\ \ A = 0, \end{cases} \tag{2}$$

where P is the predicted number of the motorcycles and A is the actual number of motorcycles. Figure 3 shows some examples of motorcycle recognition in the experiments.

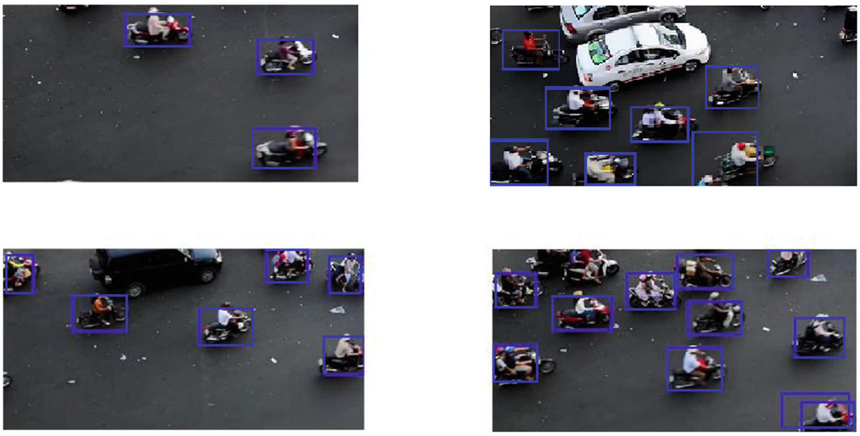


Fig. 3. Illustrative examples of motorcycle recognition

Most of the motorcycles can be judged correctly. However, when two motorcycles are too close, it may cause the two motorcycles in a bounding box and wrongly judged as one. Besides, there may be two bounding boxes with one motorcycle because the redundancy checking does not correctly remove one of them. The counting error rate in the experiments based on the formula above is 0.2 when the overlapping threshold *t* is set at 0.5.

5 Conclusions and Future Work

In this paper, the selective search algorithm and the deep learning model are integrated to recognize and count motorcycles for an image. The experimental results show the deep learning model has a good application in motorcycle counting. In the future, we will continuously improve the accuracy by testing different overlapping thresholds and use more data.

References

1. <http://www.vision.caltech.edu/html-files/archive.html>
2. <https://www.youtube.com/watch?v=Op1hdgzmhXM>
3. Bruno, A., Greco, L., La Cascia, M.: Video object recognition and modeling by SIFT matching optimization. In: Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods (2014)
4. Chung, J., Sohn, K.: Image-based learning to measure traffic density using a deep convolutional neural network. IEEE Trans. Intell. Transp. Syst. **19**(5), 1670–1675 (2018)
5. Kato, T., Ninomiya, Y., Masaki, I.: Preceding vehicle recognition based on learning from sample images. IEEE Trans. Intell. Transp. Syst. **3**(4), 252–260 (2002)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, IPS 2012, USA, vol. 1, pp. 1097–1105. Curran Associates Inc. (2012)
7. Matthews, N., An, P., Charnley, D., Harris, C.: Vehicle detection and recognition in greyscale imagery. Control Eng. Pract. **4**(4), 473–479 (1996)
8. Mukhtar, A., Tang, T.B.: Vision based motorcycle detection using HOG features. In: Proceeding of 2015 IEEE International Conference on Signal and Image Processing Applications (2015)
9. Silva, R., Aires, K., Santos, T., Abdala, K., Veras, R., Soares, A.: Automatic detection of motorcyclists without helmet. In: Proceeding of 2013 XXXIX Latin American Computing Conference (2013)
10. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. Int. J. Comput. Vision **104**(2), pp. 154–171 (2013)
11. Wen, X., Yuan, H., Song, C., Liu, W., Zhao, H.: An algorithm based on SVM ensembles for motorcycle recognition. In: Proceeding of 2007 IEEE International Conference on Vehicular Electronics and Safety (2007)
12. Zhang, Y., Pezeshki, M., Brakel, P., Zhang, S., Laurent, C., Bengio, Y., Courville, A.:Towards end-to-end speech recognition with deep convolutional neural networks. In: Interspeech, pp. 410–414 (2016)
13. Rahman, M.A., Wang, Y.: Optimizing intersection-over-union in deep neural networks for image segmentation. In: The International Symposium on Visual Computing, pp. 234–244 (2016)

Smart Transportation Technologies and Its Applications (Invited Session 04)

Algorithm of Trip Destination Estimation Based on E-Card Data – a Case Study on Taichung Urban Bus

Chao-Fu Yeh(✉) and Chi-Hua Lu

Feng-Chia University, No.100, Wenhua Rd., Xitun dist., Taichung city, Taiwan
cfyeh@fcu.edu.tw

Abstract. Smart card data contains a lot of useful information, so both of urban bus and intercity bus are built up the contactless smart card system in Taiwan. Although smart card data can offer useful information, it's still limited to how the passengers are being charged. This paper focuses on the estimation of the destination of the passenger who takes the bus twice a day and presents an algorithm to estimate the destination location for each individual boarding of the bus with a smart card. To evaluate the accuracy of destination estimation algorithm, this paper selected Taichung urban bus which uses mileage charges as a case study area. According to the results, the estimated accuracy is 63% with the 600 m tolerance distance (allowable walking distance), after considering 500 m of the value of tolerance distance error the accuracy increases to 90%.

Keywords: Smart card data · Destination estimation · Taichung urban bus

1 Introduction

Smart card data have been used in different domestic and foreign research areas because smart card data are more precise and all-round data compared with traditional travel survey. Smart card data contains a lot of useful information, e.g. date, time, bus stop, boarding and alighting stop/time, ID number of smart card and so on.

In addition to providing passengers with convenient payment methods and providing bus companies with safer and faster ticket clearing, smart card system can provide each transaction of passenger trip which can record the passenger behavior of public transportation. Thus, using smart card data can give an important for the planning and operation management of the public transportation system. Because of the advantages of the smart card, both of urban bus and intercity bus are built up the contactless smart card system in Taiwan.

Although smart card data can offer useful information, it's still limited to how the passengers are being charged, e.g. mileage charges or pay-per-segment. If the bus fare is charged through the mileage charges, the smart card database can collect the boarding and the alighting stop of the passenger. If the bus fare is charged through pay-per-segment, the smart card database can only collect the data for boarding or the alighting stop. Pay-per-segment causes an incomplete data, so the data can't be used for the transportation planning directly, like most city in Taiwan.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 311–319, 2019.
https://doi.org/10.1007/978-3-030-04585-2_38

This paper aims at estimating the destination of passengers who takes the bus and building an algorithm to estimate the destination location for bus trip using smart card with pay-per-segment. Also, the case study on our research focus on Taichung urban bus which uses smart card system with mileage charges to evaluate the accuracy of destination estimation algorithm and analyze the passenger's mode of trip.

2 Literature Review

Sohn (2017) have proposed a deep learning method for estimating origin-destination of bus trip by using the smart card data. This article was first research focusing on OD estimation of bus trip using deep learning method. This research used the smart card data for 4.8 million bus users in Seoul, Korea (more 10 million electronic ticket transactions per day), it still brought into the land use and total floor area information. The result of Jung and Sohn's research shows that the accuracy is 6% to 7% higher than the traditional analysis method.

2.1 Origin Estimation

Zhao et al. (2007) used the electronic ticket in Chicago city, USA, to combine the train electronic ticket and the bus GPS data to estimate the O-D table of train trip. Farzin (2008) used the GPS data and electronic ticket data to estimate the origin-destination of urban bus in Sao Paulo, Brazil. To solve the imperfect defect of one-ticket smart card system and analyzed the basis bus passenger flow, Yin et al. (2010) combined bus information with smart card data to find each passenger's bus route is on which direction.

Nassir et al. (2011) combined the bus schedule table, electronic ticket information and automatic passenger counter to estimate the origin-destination for bus trip. Wang et al. (2011) developed a model for predicting the trip chain of public transportation, using Oyster and iBus data to record the passenger's boarding position and vehicle, and the origin point of next trip would be estimated as a destination point of last trip.

To analysis of bus passengers' travel characteristics, Yang et al. (2013) proposed the method to identify the boarding stops. The method combined passenger's boarding time from smart card data with bus GPS data to determine passenger is on which latitude and longitude at that boarding time.

2.2 Destination Estimation

Barry et al. (2002) proposed the entry-only information for the New York subway operation combining the electronic ticket information such as the subway station data, train arrival time, user's E-card number and user's arrival time to estimate simply the OD table of train trip.

Trépanier (2007) mentioned that in most smart card automated fare collection system passenger's alighting location are not validated. To make smart card data can be applied in transit planning, present a model to estimate the destination for each passenger. The model is based on finding the shortest distance between the stop which on the route of passenger's first trip and the boarding stop of passenger's second trip.

Ma et al. (2012) used the Markov chain based Bayesian decision tree algorithm and the electronic ticket transaction data of urban buses to estimate passengers the OD point of two urban bus routes in Beijing.

Munizaga et al. (2012), mainly followed the research result of trip chain estimation method proposed by Trépanier et al. (2007), proposed an improvement method and applied this new improvement method to the estimation of trip OD in Santiago, the capital of Chile. In addition to the AVL and Geographic Information System (GIS) data, the travel location (latitude and longitude), boarding time from the transaction data and minimum allowable walking time are used to estimate the destination of last trip. This method can also handle the round-trip (original route) round-trip when the bus route is used. Ma et al. (2013) used density-based spatial clustering of application with noise algorithm (DBSCAN) and the electronic ticket transaction data of the urban bus to estimate two urban routes in Beijing.

If the distance between two stops is smaller than the tolerance distance then labels the stop as estimated alighting stop. If the distance between two stops is bigger than the tolerance distance then applies Kernel density estimation to calculate the probability of which stop is the most possible estimated alighting stop (He et al. 2015).

2.3 Origin-Destination Estimation

Ro (2008) used smart card data of Taipei city to propose the method to produce the O-D table of bus line. The study focused on regular passengers which the number of transactions is more than two-thirds of the day of the data and they have the transaction records both on morning and evening peak hours.

Gao et al. (2015) proposed a method to calculate the route O-D table. The method estimated the boarding stop of the passengers by matching the smart card data with the bus stop data and estimated the alighting stop by dividing the number of records of the day into three categories. When there is only one record of the passenger, more records need to be involved. When there are two records of the passenger, if they're from the same route, the boarding stop of the next trip can be labeled as the estimated alighting stop.

From the above-mentioned retrospective literature, it can be found that the research related to electronic ticket has increased greatly in recent years. The important contents of the relevant documents for the estimation of OD table by electronic ticket can presented several estimation methods for origin-destination point of trip.

3 Methodology of Destination Estimation

3.1 Smart Card Data

To reduce the estimation errors that caused by the mistakes of original data, data preprocessing is an important step. First, defined the meanings of 24 fields in the smart card database, including ID number, date, route number, boarding stop, alighting stop, ticket type, boarding time, alighting time and so on. Second, selected the fields that be applied on destination estimation and validation, including date, direction, ID number, route number, boarding time, boarding stop number and alighting stop number. Third, excluded the following data:

- Exclude the data when passenger's ticket type belongs to token. In this case, we can't be sure that the same ID number of the token is from the same passenger.
- Exclude the data when there is no record of passenger's alighting stop. In this case, we can't determine that the destination estimation is correct or not.
- Exclude the data when passenger's alighting and boarding stops are the same. In this case, this kind of travel behavior is unusual.
- Exclude the data when there is an error that can't be fixed.

3.2 Destination Estimation Algorithm

This section explains five factors that have considered on destination estimation algorithm, including boarding time, destination assumption, direction, tolerance distance and value of tolerance distance error.

- Boarding time: Sorting the passenger's trip by boarding time, to know which trip happened first and which happened next.
- Destination assumption: Based on trip-chaining theory, passenger's alighting stop is usually near the next boarding stop and the last trip of the day, the alighting stop is usually close to the first trip's boarding stop.
- Direction: The same bus route has outward and return two directions, considerate of which way the bus is on to select the correct longitude and latitude of the bus stop. Also, the estimated alighting stop needs to match the direction of the bus, as shown in Figs. 1 and 2.
- Tolerance distance: Tolerance distance is to determine if the distance between the boarding stop and the estimated alighting stop is acceptable.
- Value of tolerance distance error: Value of tolerance distance error is the value of acceptable walking distance because considerate of the distance between two bus stops are close in Taiwan. It's possible that passenger may choose other bus stop as alighting stop around the closest estimated alighting stop.

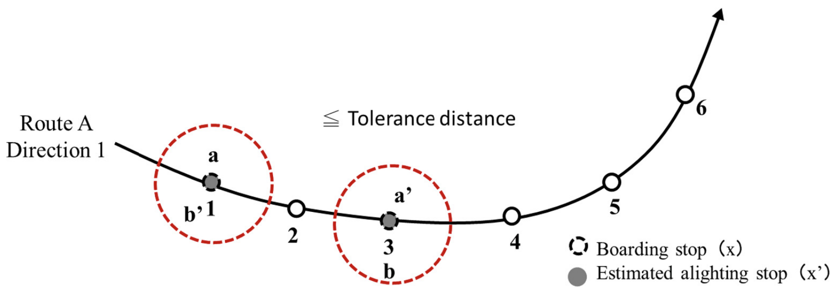


Fig. 1. Before consider the bus direction

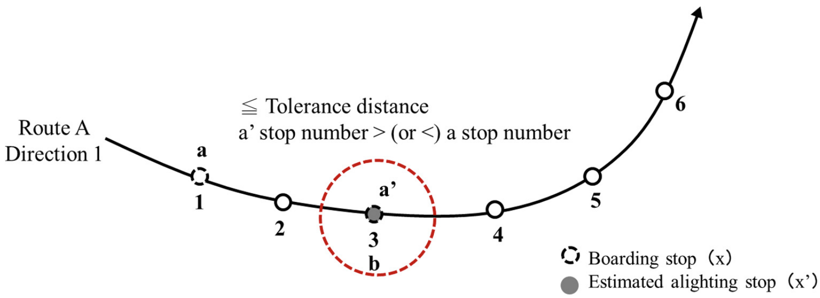


Fig. 2. After consider the bus direction

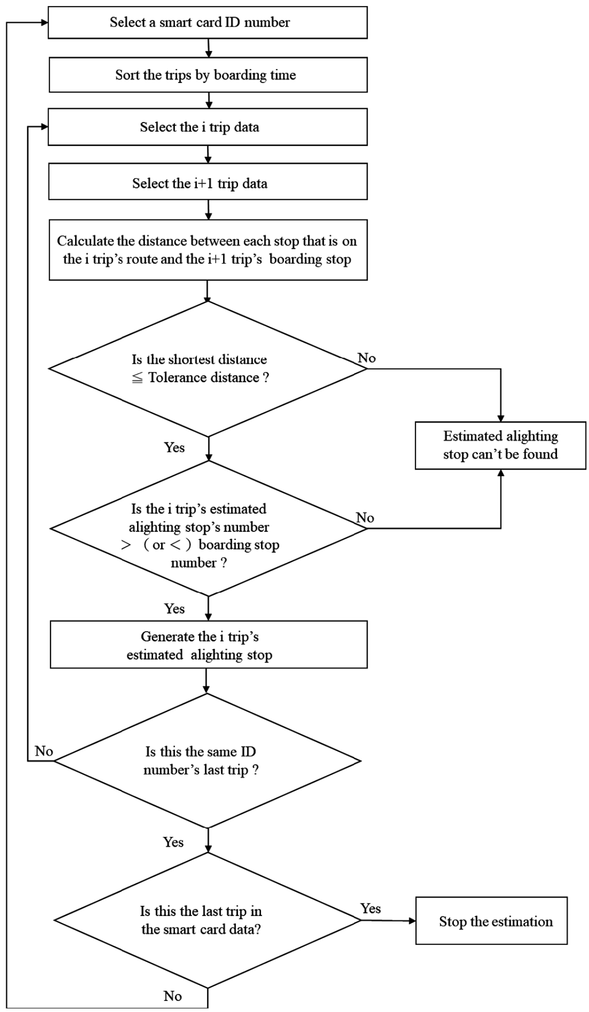


Fig. 3. Destination estimation algorithm

As shown in Fig. 3 above, based on these five factors above, the steps of finding passenger's estimated alighting stop of each trip is as following:

1. Select a smart card ID number on the same date and sort the passenger's trips by boarding time.
2. Select the i trip and the i + 1 trip data, and calculate the distance between each stop that is on the i trip's route and the i + 1 trip's boarding stop.
3. If the shortest distance between two stops is smaller than the tolerance distance then goes to the next step. If the shortest distance between two stops is bigger than the tolerance distance then this trip's estimated alighting stop can't be found.
4. If the estimated alighting stop number match how the stop number is defined on that route then label it as estimated alighting stop; otherwise the trip's estimated alighting stop can't be found.
5. Continue the process until is the last trip in the smart card data.

4 Case Study on Taichung Urban Bus

Taichung city uses mileage charges which includes data on both boarding and alighting stop of the passenger. To evaluate the accuracy of destination estimation algorithm, cover up the actual alighting stop and estimate the alighting stop using destination estimation algorithm. Then, compare the estimated alighting stop with the actual alighting stop to see if the two stops are the same. Select a week from the smart card data randomly. According to the statistics, 48% to 54% passengers take the bus once a day and 33% to 38% passengers take the bus twice a day from 12/1/2014 to 12/7/2014. If the passenger only has one record in a day then in this situation the estimated alighting stop can't be found through a day. Hence, our study focuses on the passenger who takes the bus twice in a day.

4.1 Destination Estimation and Validation

Select 12/1/2014 smart card data as sample, there are 133,722 trips that the passengers take the bus twice in a day. After excluding the following trips: ticket type belongs to token, no record of alighting stop, alighting and boarding stops are the same and an error that can't be fixed. There are 107,808(81%) trips left that can be used in the further estimation.

When the tolerance distance is bigger than 600 meters, the number of trips that can be estimated haven't had much different. When the tolerance distance is 1000 meters, the accuracy is around 62%. When the tolerance distance is lower to 100 meters, the accuracy increases to 66%. These show that the accuracy decreases with the tolerance distance increases. If we consider the value of tolerance distance error, select tolerance distance 600 meters as sample, the accuracy increases with the value of tolerance distance error increases. This means some passengers choose the alighting stop which near the estimated alighting stop instead of the closest stop.

4.2 Passenger's Mode of Trip Analysis

Based on trip-chaining theory, passenger's alighting stop is usually near the next boarding stop and the last alighting stop of the day is usually close to the first trip's boarding stop. To classify the passenger's mode of trip, calculate the distance between the first trip's boarding stop and the second trip's alighting stop. Also, calculate the distance between the first trip's alighting stop and the second trip's boarding stop, to determine if these two stops are smaller or bigger than the tolerance distance.

After calculate the distance between the stops, classify the mode of trip by route and direction to see if the two trips are from the same route or direction. The mode of trip can be classified in 12 categories. According to the statistics, most of the trips that can be analyzed and estimated are from the mode 2-2, 3-2 and 3-4, as shown in Fig. 4. The number of trips difference between analysis and estimation are due to the distance between the boarding stop and estimated alighting stop is bigger than tolerance distance or the estimated alighting stop doesn't match the direction of the bus route. In these 12 categories, the mode 1-4 and 2-2 have the highest accuracy of 71%. The mode 3-2 has the third highest accuracy of 65%, as shown in Fig. 5.

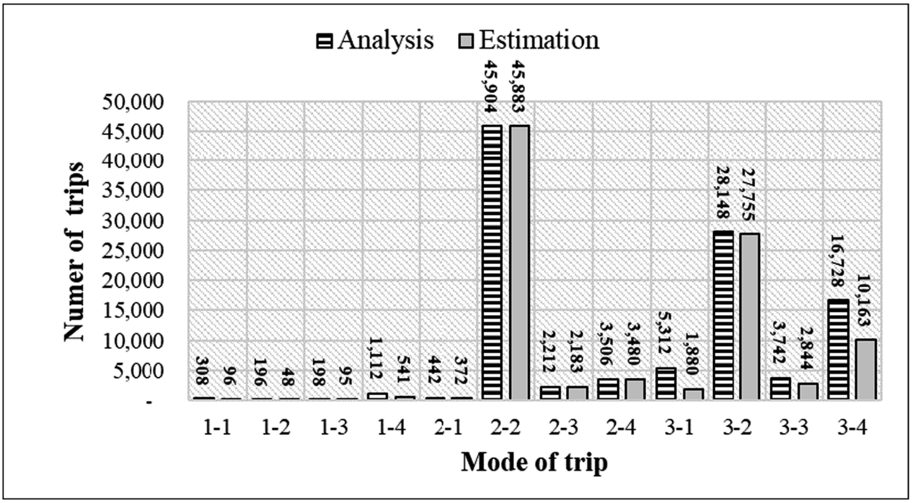


Fig. 4. Number of trips under different mode of trip

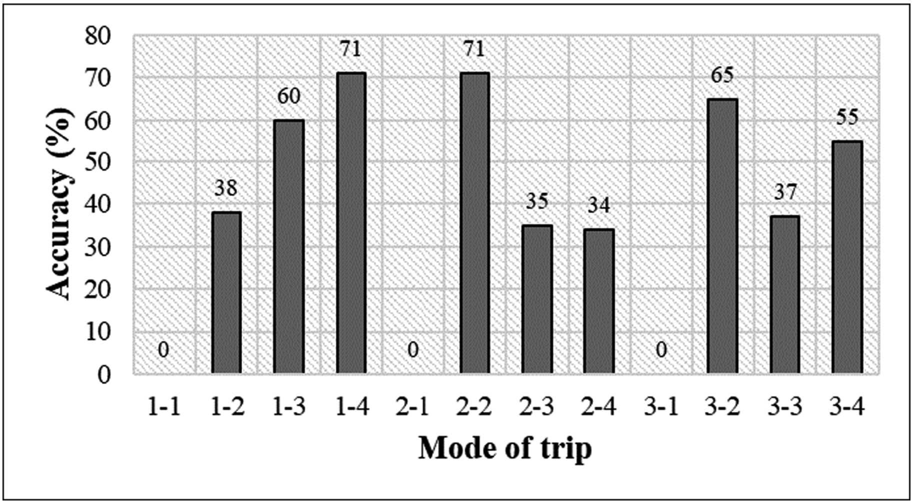


Fig. 5. The accuracy under different mode of trip

5 Conclusions

Smart card data contains a lot of useful information, but its data is still limited to how the passengers are being charged, e.g. mileage charges or pay-per-segment. When the bus fare is charged through pay-per-segment, it will cause an incomplete data. To deal with this issue, our study focuses on the estimation of the destination of the passenger who takes the bus twice a day. The algorithm to estimate the destination location for each individual boarding of the bus with a smart card relates to five factors, including the boarding time, destination assumption, direction, tolerance distance and value of tolerance error.

To evaluate the accuracy of destination estimation algorithm, select Taichung city as a case study area. The smaller the tolerance distance is, the fewer trips that can be estimated. When the tolerance distance is 1000 meters, the accuracy is around 62% and when the tolerance distance is lower to 100 meters, the accuracy increases to 66%. These show that the accuracy decreases with the tolerance distance increases and the trips that can be estimated and the accuracy are in the reverse relation under the same tolerance distance. After considering the value of tolerance distance error, the accuracy increases with the value of tolerance distance error increases. For example, the original accuracy is 63% of the 600 meters tolerance distance, after considering 500 meters of value of tolerance distance error the accuracy increases to 90%. This means 27% of trips passengers choose the alighting stop which near the estimated alighting stop instead of the closest stop. After classifying the mode of trip in 12 categories, most of the trips that can be analyzed and estimated are from the mode 2-2, 3-2 and 3-4. The mode 1-4 and 2-2 have the highest accuracy of 71% and the mode 3-2 has the third highest accuracy of 65%.

References

Barry, J., Newhouser, R., Rahbee, A., Sayeda, S.: Origin and destination estimation in New York City with automated fare system data. Transp. Res. Rec. **1817**, 183–187 (2002)

Farzin, J.: Constructing an automated bus origin-destination matrix using fare card and GPS Data in São Paulo, Brazil. Presented at the 87th TRB Annual Conference, Washington, DC (2008)

Gao, L.X., Li, G.Y, Hu, J.H, Liang, J.X.: A calculation method of OD matrix in multi-modal transit network based on traffic big data. In: International Conference on Transportation Information and Safety, pp. 295–298 (2015)

He, L., Nassir, N., Trépanier, M., Hickman, M.: Validating and calibrating a destination estimation algorithm for public transport smart card fare collection systems. No. CIRRELT-2015-52 (2015)

Ma, X.L., Wang, Y.H., Chen, F., Liu, J.F.: Transit smart card data mining for passenger origin information extraction. J. Zhejiang Univ. Sci. C **13**(10), 750–760 (2012)

Munizaga, M.A., Palma, C.: Estimation of a disaggregate multimodel public transportation origin-destination matrix from passive smartcard data from Santiago, Chile. Transp. Res. Part C **24**, 9–18 (2012)

Nassir, N., Khani, A., Lee, S.G., Noh, H., Hickman, M.: Transit stop-level origin-destination estimation through use of transit schedule and automated data collection system. Transp. Res. Rec. **2263**(2011), 140–150 (2011)

Ro, W.Y.: Using Taipei Easycard transation data to explore the O-D table of bus passengers. Thesis, Tamkang University (2008)

Trépanier, M., Chapleau, R., Tranchant, N.: Individual trip destination estimation in a transit smart card automated fare collection system. J. Intell. Transp. Syst. Technol. Plann. Oper. **11**, 1–14 (2007)

Wang, W., Attanucci, J.P., Wilson, N.H.: Bus passenger origin-destination estimation and related analyses using automated data collection systems. J. Public Transp. **14**, 131–150 (2011)

Yang, D.Y., Chen, J.: Identifying boarding stops of bus passengers with smart cards based on intelligent dispatching data. J. Transp. Syst. Eng. Inf. Technol. **13**(1), 76–80 (2013)

Yin, C.Y., Chen, Y.Y., Chen, S.H.: Bus station passenger matching method based on cluster analysis method. J. Transp. Inf. Saf. **28**(3), 21–24 (2010)

Zhao, J., Rahbee, A., Wilson, N.H.M.: Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. Comput.-Aided Civil Infrastr. Eng. **22**(5), 376–387 (2007)

Vehicle Miles Traveled Estimation Based on Taxi GPS Data: A Case Study in Nanjing, China

Guopeng Zhang[1], Liang Xia[1], Wenbo Fan[1,2], Guangrong Chen[3], Yingfei Fan[1], and Xinguo Jiang[1,2(✉)]

[1] School of Transportation and Logistics, Southwest Jiaotong University, West Park, High-Tech District, Chengdu 611756, China
ejiang@gmail.com
[2] National United Engineering Laboratory of Integrated and Intelligent Transportation, West Park, High-Tech District, Chengdu 611756, China
[3] Shenzhen Urban Transport Planning and Design Institute Co., Ltd., 1700 Tianfu Road, High-Tech Zone, Chengdu 610041, China

Abstract. Vehicle miles traveled (VMT) is one of the basic metrics with applications in almost every aspect of traffic engineering and safety. Traditionally, there are many methods available to estimate the VMT, e.g., traffic volume counts, odometer readings, household questionnaire surveys, and fuel sale data. However, limitations of these techniques include high data-collection costs, low accuracy, and failure to estimate the VMT at the finely disaggregated levels. The study aims to develop an approach to estimate the VMT of the operating taxis according to the readily available GPS data, which bears the advantage of being accurate, all-weather, continuous, and less expensive. The study consists of: (1) identifying and removing the abnormal data with the potential issues from the GPS database; (2) computing the total VMT by summing the distances of every two sequential GPS points; and (3) validating the estimated VMT data from the practical and statistical perspectives (by error analysis and two-sample Kolmogorov-Smirnov test). The GPS data of operating taxis in Nanjing, China are used for the empirical study. The study serves to provide a feasible solution for the VMT estimation.

Keywords: VMT · GPS technique · Operating taxi · Travel distance Validation

1 Introduction

During the last decades, it has seen the surging popularity of vehicle miles traveled (VMT) in the field of traffic engineering, which is defined as the total miles traveled by all vehicles within a particular area and timeframe. VMT is an important indicator of traffic exposure. It has been broadly applied in the field of traffic engineering, ranging from planning, operations, and management to safety.

In the field of traffic safety, VMT is the most widely used measurement of the crash exposure to quantify the hazardous environments for the crash occurrence [1]. There are a

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 320–327, 2019.
https://doi.org/10.1007/978-3-030-04585-2_39

large number of VMT-based safety assessments in the traditional researches and applications [2, 3]. Nonetheless, there are two major limitations associated with the VMT applications: (1) the assumption of "linear conjecture" between crash frequency and VMT has been constantly challenged over the years [4]. It is based upon the fact that the high VMT is typically accumulated on the relatively higher functional roadways with limited access and medians, while the low VMT is accumulated on the lower level of roadways which are undivided [5]; and (2) it is virtually impossible to obtain the VMT data disaggregated by the particular characteristics (e.g., specific spatial and temporal variables) [6].

Reliable estimates of VMT are important for a number of traffic applications. Traditionally, a variety of methods have been developed for the VMT estimation, which can be generally classified into two major categories: (1) traffic-count-based (TC-based) methods and (2) non-traffic-count-based (NTC-based) methods [7]. Comparatively, the TC-based methods produce more reliable estimates because they are directly based upon the actual traffic count. The NTC-based methods mainly utilize the data sources such as odometer readings [8], and household questionnaire survey [9].

Specifically, the readily available Global Position System (GPS) data pose the great potentials for the VMT estimation. Given the predominance, there have been few studies contributing to the GPS-based VMT calculation. One of the advantages of the GPS technique is that it can provide relatively up-to-date information on the moving vehicles. Moreover, using the GPS to track vehicle movements is more cost-effective compared to other methods. Recently, the GPS-based VMT estimation were studied, such as Zhang and He [10]. Typically, the basic idea of testing the validity of method is to compare the estimates with the actual information (e.g., recalled trips for the travel day [11] or the computer assisted telephone interviewing [12]). However, the collected information could not directly reflect the value of the actual VMT. In addition, it is great importance to conduct the data manipulation prior to the VMT calculation in order to ensure the accuracy of estimation results [12].

Thus, the paper proposes an innovative GPS-based method for the taxi VMT estimation, which includes three major steps: (1) manipulating the raw GPS data; (2) computing the total taxi VMT; and (3) verifying the validity by comparing it with the corresponding exposure truth. A case study is conducted with the taxi GPS data in Nanjing (China).

2 Methodology

2.1 Data Processing Procedure

The recording process of GPS data may be accompanied with data errors, possibly due to the power outage, the human factors (e.g., switch off by humans), device fault, and/or signal problems. The following types of GPS records can be identified as the abnormal data: (1) GPS records falling outside the study area; (2) The GPS points that deviate from the actual trajectory of vehicular movement; (3) Two consecutive GPS records with distance beyond the upper limit of the running vehicle; (4) Two consecutive GPS records with very short distance (e.g., 10 m); and (5) The data that are not recorded in the database.

Multiple steps are developed to eliminate the above-stated abnormal data from the GPS dataset. First, we screen out the first two types of GPS records. Second, we break down the 3rd type of GPS records into two trips. Third, for the 4th type of data, we keep the first GPS records and remove the others. Last, we disregard the 5th type of GPS records.

2.2 VMT Estimation

The basic idea of measuring the total VMT is to sum all the travel distances between every two sequential GPS points. The travel distance between GPS point i and point $(i+1)$ can be computed with the approach proposed by Qi [13]. Then, the total VMT is the sum of $D_{i,i+1}$ of all the operating vehicles in the roadway network. It can be expressed as:

$$VMT = \sum\nolimits_{k=1}^{K} \sum\nolimits_{j=1}^{J} \sum\nolimits_{i=1}^{I} D_{i,i+1,j,k} \tag{1}$$

where $D_{i,i+1,j,k}$ is the distance traveled by vehicle k during the time period i in day j(detailly described in Qi [13]), K is the total number of vehicles, J is the number of days, and I is the number of the total time periods in one day. Note that the above VMT estimation can be conducted for any specific time of day, e.g., peak hours.

2.3 Validity Test

The validity of the GPS-based VMT estimation should be conducted by comparing the estimates with the exposure truth (the true VMT), which can be obtained by measuring the length of the actual trajectory of the vehicular movements (i.e., through, turning, and U-turn) on Google Earth Map. The following rules are adopted to determine the moving trajectory of vehicles: (1) Through movement takes place via the shortest path on the straight road sections; (2) Curvilinear movement occurs on ramps of the curvy road sections; and (3) Turning movements (i.e., left-turning, right-turning, and U-turn).

The process of measuring exposure truth includes three steps: (1) locating the GPS points on the Google Earth; (2) identifying the travel path of the vehicles; and (3) measuring the actual distance of the vehicles traveled on the Google Map. The total VMT equals the sum of distances between every two sequential GPS points, as follows:

$$VMT' = \sum\nolimits_{k=1}^{K} \sum\nolimits_{j=1}^{J} \sum\nolimits_{i=1}^{I} D'_{i,i+1,j,k} \tag{2}$$

where VMT' is the total VMT of the target vehicle, $D'_{i,i+1,j,k}$ is the actual distance traveled by the vehicle k during the time period i in day j.

The comparison between the estimated VMT and true VMT can be conducted from both practical and statistical perspectives. For the former, the authors introduce the absolute error (AE) and relative error (RE) to show the estimation error of the travel distance during a time period.

$$AE_i = \left| D_{i,i+1,j,k} - D'_{i,i+1,j,k} \right| \tag{3}$$

$$RE_i = \frac{\left| D_{i,i+1,j,k} - D'_{i,i+1,j,k} \right|}{D'_{i,i+1,j,k}} \times 100\% \tag{4}$$

The total relative error (TRE) is introduced to describe the estimation error of the total VMT. It is expressed as follows:

$$TRE = \frac{\left| VMT - VMT' \right|}{VMT'} \times 100\% \tag{5}$$

An empirical value of 4% [14] is selected as the threshold value to identify the practical difference of two datasets. In order to describe the distribution of RE_i, the authors further employ the accumulated relative error (ARE) and accumulated percentage of relative error (APRE), which are expressed as:

$$ARE_i = \frac{\sum_{j=1}^{i} AE_j}{VMT'} \times 100\% \tag{6}$$

$$APRE_i = \frac{ARE_i}{ARE_I} \times 100\% = \frac{\sum_{j=1}^{i} AE_j}{\sum_{j=1}^{I} AE_j} \times 100\% \tag{7}$$

From the statistical view, the authors employ two-sample Kolmogorov-Smirnov (K-S) test to compare the estimated VMT and true VMT and consequently assess the accuracy of the proposed method. The K-S test is a non-parametric statistical test that is commonly used to compare the observed and expected data [14].

3 Case Study

The authors collected the taxi GPS data in the city of Nanjing, China on Sep. 1, 2010. Overall, 1849 taxies are included in the database with 4.8 million data records. The city of Nanjing with the area of 8416.3 square kilometers is chosen as the study area (from 118°22’ to 119°14’ eastern longitude, and from 31°14’ to 32°37” northern latitude).

3.1 VMT Estimation

Figure 1 illustrates the distribution of estimated VMT and passenger load factor (PLF). The total VMT estimate is 472.7 million km per day. Most of them are accumulated during the daytime instead of the nighttime. During the time period 12am-1 pm, the VMT reaches the maximum value, while the minimum value of VMT occurs during the time period from 4am to 5am when the travel demands are low.

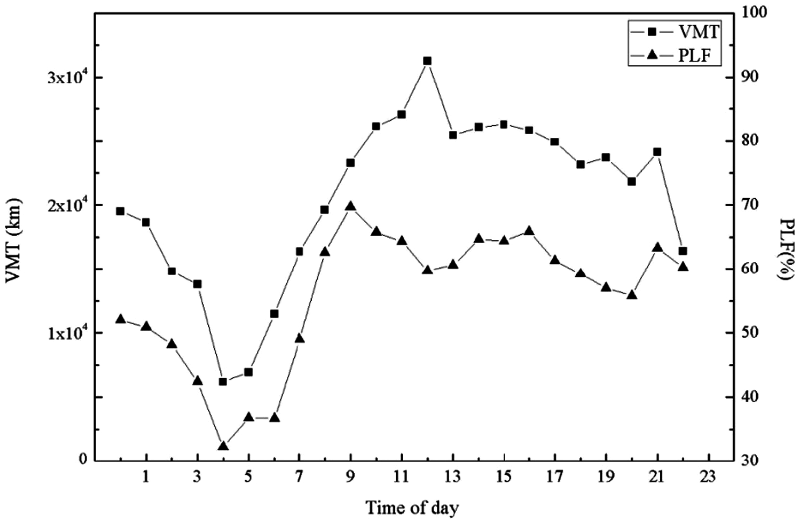


Fig. 1. The distributions of estimated VMT and the PLF by time of day

It is also observed that the PLF at daytime is much greater than that at nighttime. The minimum value also occurs between 4am and 5am. However, the maximum value of PLF is in the time period from 9am to 10am; and then it demonstrates the decreasing trend until 12 pm. The observation indicates that the VMT for taxis with passengers on board has a different distribution pattern compared to the total VMT. It can be explained by the fact that during the non-working hours (e.g., at noon) the relatively less travel demand leads to a low PLF; in the meanwhile, most of the taxi vehicles operate in the roadway network at a relatively high speed (due to the relieved traffic congestion during the non-peak hours), which consequently produces a higher VMT.

3.2 Error Analysis

To evaluate the accuracy of the VMT estimation, the authors select a random set of GPS data from the database to conduct the comparison between the VMT estimates and the corresponding true VMT. As shown in Table 1, the estimated VMT generally approximates the VMT truth (316.5 km vs. 329.8 km). The TRE of the estimated result is 4.0%, implying an acceptable accuracy. Moreover, the two-sample K-S test yields the p-value of 0.129, suggesting a consistent distribution between the estimated VMT and the exposure truth. Therefore, it suffices to demonstrate the validity and accuracy of the GPS-based VMT estimation method.

Table 1. Comparison of VMT estimated by GPS-based method and true VMT

Parameters	GPS-based VMT	True VMT	|Difference|%
VMT (km)	316.5	329.8	4.0%
Mean of $D_{i,i+1}$ (m)	199.1	207.4	4.0%
Standard deviation (m)	153.5	165.3	7.1%
Two-sample K-S test	D = 0.041, p-value = 0.129		
Total relative error	4.0%		

To further explore the estimation error, the authors categorize the travel trajectories into 3 groups: (i) straight tracks (ST), (ii) curve tracks without turning (CT), and (iii) turning tracks (TT) according to the actual trajectory of the moving taxies on the Google Earth Map. Statistical analysis is conducted for the above tracks of the GPS records. Figures 2 and 3 show the distributions of RE and APRE and accumulated travel miles (ATM) for various travel tracks, respectively. The information displayed in Fig. 2 indicates that all the REs in the scenario "ST" are smaller than TRE. For the scenario "CT," most of the ERs fall below the empirical threshold value (4%). However, it is pinpointed that the ERs under the scenario "TT" are much more dispersed than others (approximately 90% of the ERs > 4.0%). The information shown in Fig. 3 indicates that a majority of the operating taxies' trajectories are composed of straight movements (nearly 66%), which leads to a relatively high ATM (57.0%). The corresponding APRE accounts for only 4.6% of the total APRE. Therefore, it is safe to argue that the VMT estimation based on the GPS is highly accurate for the straight trajectories. It is because the calculated VMT between consecutive GPS points is the linear distance, which is consistent with the straight travel trajectories. Under the scenario "CT," the curve trajectories account for 12.8% of the total APRE and 22.9% of ATM. It means that the accuracy of VMT estimates is slightly influenced by the presence of the curvy section of roadways. While for the scenario "TT", the APRE accounts for 82.5% of the total APRE, which is inconsistent with the corresponding ATM percentage (20.1% only). Obviously, the error of the total VMT estimation is mainly caused by the turning movement of taxi vehicles. Fortunately, the proportion of "TT" (nearly 19%) is relatively smaller than other two types of travel trajectories, so its contribution to the estimation error is within the acceptable tolerance.

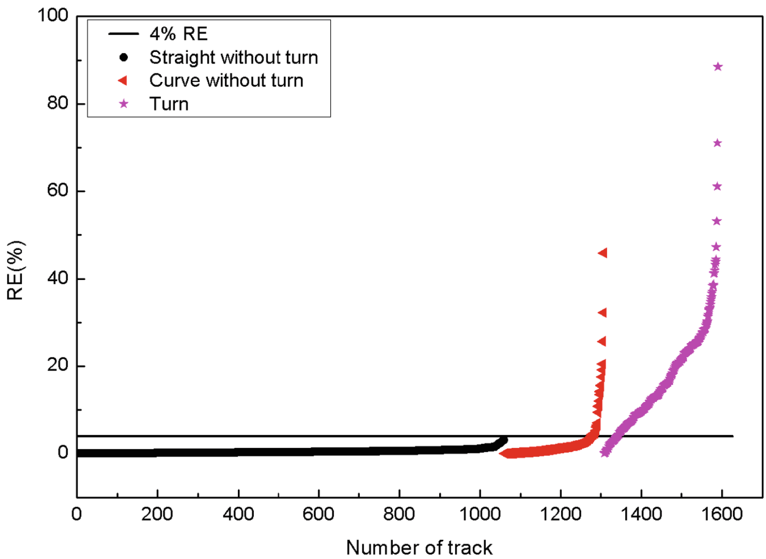


Fig. 2. Distributions of travel distance ER for various travel tracks

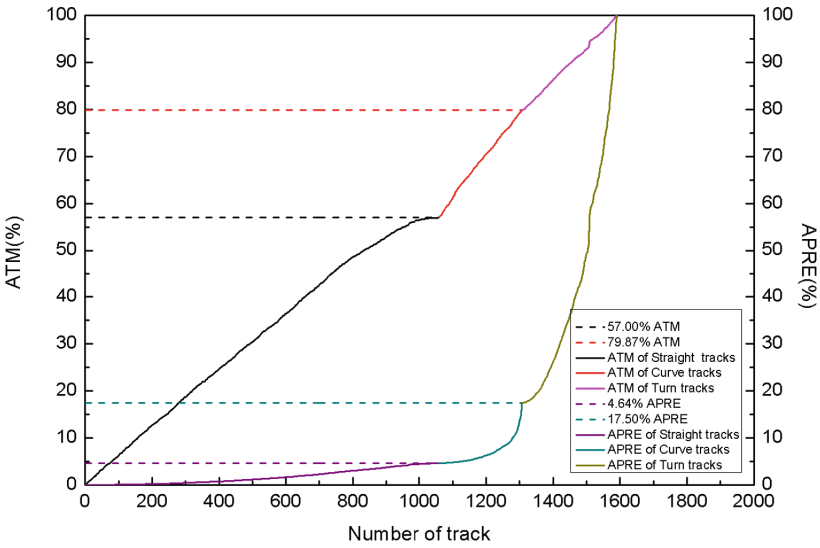


Fig. 3. Distributions of ATM and APRE for various travel tracks

4 Discussions

The study proposes a GPS-based method to estimate the taxi VMT. The taxi GPS data in Nanjing (China) are utilized to verify the validity of the method. The results consistently demonstrate that the proposed method is able to produce a relatively accurate VMT estimate.

Except for the accurate estimating results, there are obvious advantages of the GPS-based method over the conventional VMT estimation. Since the GPS information of moving vehicles are usually generated in the real time, the proposed method can provide relatively up-to-date VMT as opposed to other traditional ones (e.g., household survey and fuel sale survey). Moreover, the GPS-based VMT estimation is convenient and economical, without the efforts of collecting extra information such as the traffic volume and the segment length of roads. In addition, the GPS-based method also poses the great advantage to calculate the VMT disaggregated by particular characteristics of interest, such as time of day, the state of passenger occupancy, and specific spatial characteristics (e.g., functional roadway classifications). Especially, the information is of great value to improve the management of operating vehicles for the taxi company.

Limitations of the study should be mentioned. Inaccuracy is somewhat associated with the estimation process, particularly for calculating the travel distance of the turning movement. Consequently, estimation error exists in the VMT along the curvy sections of roads. The calculated VMT is based on the linear distance between GPS points, while the actual vehicular trajectory is always a curvy line, which may lead to the discrepancy between the estimated VMT and actual VMT. Future work will be directed to focus on the VMT estimation of vehicles with turning movements so as to improve the accuracy of the overall VMT estimation. Notwithstanding, the proposed technique to measure the taxi VMT can be conveniently extended to other public vehicles equipped with the GPS device.

Acknowledgements. The research is supported by the National Science Foundation of China (NSFC-71771191) and Doctoral Innovation Fund Program of Southwest Jiaotong University.

References

1. Jiang, X., Lyles, R.W., Guo, R.: A comprehensive review on the quasi-induced exposure technique. Accid. Anal. Prev. **65**, 36–46 (2014)
2. Derrig, R.A., Segui-Gomez, M., Abtahi, A., Liu, L.L.: The effect of population safety belt usage rates on motor vehicle-related fatalities. Accid. Anal. Prev. **34**(1), 101–110 (2002)
3. Chi, G., Cosby, A.G., Quddus, M.A., Gilbert, P.A., Levinson, D.: Gasoline prices and traffic safety in Mississippi. J. Saf. Res. **41**(6), 493–500 (2010)
4. Janke, M.K.: Accidents, mileage, and the exaggeration of risk. Accid. Anal. Prev. **23**(2), 183–188 (1991)
5. Chandraratna, S., Stamatiadis, N.: Quasi-induced exposure method: evaluation of not-at-fault assumption. Accid. Anal. Prev. **41**(2), 308–313 (2009)
6. Lyles, R.W., Stamatiadis, P., Lighthizer, D.R.: Quasi-induced exposure revisited. Accid. Anal. Prev. **23**(4), 275–285 (1991)
7. Kumapley, R., Fricker, J.: Review of methods for estimating vehicle miles traveled. Transp. Res. Rec. **1551**, 59–66 (1996)
8. Fox, J., King, F.: Using odometer readings as panel data to estimate historical vehicle kilometers travelled for light-duty vehicles in metro Vancouver. In: 94th Annual Meeting Transportation Research Board (2015)
9. Jiang, X., Qiu, Y., Lyles, R.: US national household travel survey used to validate exposure estimates by the quasi-induced exposure technique. Transp. Res. Rec. **2237**, 152–159 (2011)
10. Zhang, L., He, X.: Feasibility and advantages of estimating local road vehicle miles traveled on basis of global positioning system travel data. Transp. Res. Rec. **2399**, 94–102 (2013)
11. Wagner, D.P.: Lexington area travel data collection test: GPS for personal travel surveys, Office of Highway Policy Information and Office of Technology Applications, Federal Highway Administration, Battelle Transport Division, Columbus (1997)
12. Wolf, J., Oliveira, M., Thompson, M.: The impact of trip underreporting on VMT and travel time estimates: preliminary findings from the California statewide household travel survey GPS study. Transp. Res. Rec. **1854**, 189–198 (2003)
13. Qi, L.: Research and application on taxi traffic running performance based on GPS DATA. Harbin Institute of Technology, Harbin, China (2013). (Master dissertation)
14. Jiang, X., Lyles, R.W.: A review of the validity of the underlying assumptions of quasi-induced exposure. Accid. Anal. Prev. **42**(4), 1352–1358 (2010)

Travel Time Prediction Based on Missing Data Compensation

Tang-Hsien Chang[1(✉)], Yi-Ru Li[2], Cheng-Hong Fu[1], Yao-Bin Liu[1], and Shu-Min Yang[1]

[1] School of Transportation, Fujian University of Technology, Fuzhou, China
thchang@fjut.edu.cn, fch407@126.com, 2627789795@qq.com
[2] Department of Civil Engineering, National Taiwan University, Taipei, Taiwan
thchang@ntu.edu.tw

Abstract. Data Missing is a common issue in traffic data processing. This study proposes pretime interpolation and realtime extrapolation for missed data compensation. Herein, cubic spline curve interpolation, linear interpolation, and near value method are applied for pretime operation. Mean value method, Fourier transform method, α-β-γ filter with mean value and α-β-γ filter with Fourier transform method are involved for realtime operation. According to the experiments, the practices result the linear interpolation is shown of the best way for pretime works. About realtime extrapolation, the proposed four methods present without significant difference. The paper finally conducts travel time prediction via the compensated database revealing good performance.

Keywords: Intelligent Transportation Systems · Missing data compensation
Travel time prediction · Kalman filter · Fourier transformation
α-β-γ filter

1 Introduction

Traffic information is the fundamental element of building Intelligent Transportation Systems (ITS). While data collection, traffic detectors will be affected by many uncertain factors, such that causing data bias or missing. This implies the negative influence of ITS implementation. Therefore, how to compensate or repair the data is an essential issue in ITS practice.

In past decade, there are many contributions related to this theme. Roughly, there can be classified into pretime compensation and realtime compensation. Pretime compensation includes time series method, gray relationship method, interpolation/extrapolation method etc. Lin [1] applied autoregression moving average (ARIMA) to interpolate the loss data got from vehicle detectors. The MAPE (Mean Absolute Percentage Error) is about 20%. Kuo [2] constructed gray residual model to repair the missing data. Wang [3] studied the adaption by the nearest value method, linear interpolation method and cubic spline curve method for different loss rates.

Corresponding to realtime compensation, Artificial Neural Network (ANN), Kalman filter (KF), fast Fourier Transform (FFT) are common used. Ying [4] studied VNNTF (Volterra neural network traffic flow model) for traffic flow prediction, which

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 328–336, 2019.
https://doi.org/10.1007/978-3-030-04585-2_40

has better efficient than Backward Propagation Neural Network (BPNN). Peeta and Anastassopoulos [5] took Fourier Transform to develop a robust compensation, which repairs the loss data applying for traffic control parameters. Chang and Huang [6] compared historical average method, Fourier Transform method, α-β-γ filter with historical average method, α-β-γ filter with Fourier Transform method for data compensation. It concludes that, while losing rate is less than 20%, α-β-γ filter with Fourier Transform reveals significant efficiency. This can work in realtime/online practice.

On the previous reviews of references, providing a complete data is the first important task in ITS implementation. This paper generalizes and proposes several pretime methods to adapt different conditions. As for realtime compensation, it definitely requires computing speed and reliable result. This paper synthesizes the characteristics, that Kalman filter processing realtime short term prediction and cooperative with Fourier Transform dealing with historical data for long term prediction is reasonable.

2 Methodology and Modelling

2.1 Missing Data Compensation

Data missing may be caused from traffic detectors' interior or exterior problems. Its operational timing for compensation or reparation can be classified into pretime method and realtime method. In general, a pretime method is for interpolation and a realtime method is for extrapolation.

A. Pretime method
   In case of a road section, the pretime compensation is triggered at a given/fixed time of a day. Check the missing data, and using the following methods for interpolation.
(1) The nearest value method:
   The way is taking the average of several nearest data for interpolation, as the example,

$$X_{i+2} = \frac{X_i + X_{i+1} + X_{i+3} + X_{i+4}}{4} \tag{1}$$

(2) Linear interpolation method:
   By a linear line through two known data, easily evaluate a certain point on the line. Given two points, (x_1, y_1) and (x_2, y_2), a point (x, y) on the line is obtained,

$$y = \frac{y_1 - y_0}{x_1 - x_0}(x - x_0) + y_0 \tag{2}$$

(3) Cubic spline curve method:
   This method applies sinusoid function to approaching the three degrees of curve. Herein, the following equation is to compute the point at time t by the Catmull-Rom curve [7] (a cubic spline curve) with four points. t is the interval of time inside the four points.

$$p(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -3 & 3 & -2 & -1 \\ 2 & -2 & 1 & 1 \end{bmatrix} \begin{bmatrix} p_i & p_{i+1} & \frac{p_{i+1}-p_{i-1}}{2} & \frac{p_{i+2}-p_i}{2} \end{bmatrix}^T \tag{3}$$

B. Realtime compensation

Making a prediction of travel time, in general, is to estimate the reliable value on the data which were obtained in advanced several time steps. If there exists a missing data, it should perform realtime compensation to fill in or adjust, and satisfying the need of data completeness. They are four methods compared each other hereafter.

(1) Average value method:

When there data missing occurs, compute the mean of those data which on the same period in the historical database. However, the way should obey the theorem of central limit.

$$I_t = (\sum_{i=1}^{N} iT_t)/N \tag{4}$$

in which, I_t denotes the compensated value at time t. N is the number of days. T_t is the travel time at time t.

(2) Fourier Transform:

Any cycle signal, $f(t)$, can be expressed by Euler's Equation, which is an exponential function with a sinusoid form. Via Fourier transformation, traffic parameters in term of time can be smoothed on demand against noise. The process is as below: Firstly, take the data series, $x(t)$, its time domain function $f(k\Delta t)$, can be transformed into frequency domain function, $F\left(\frac{n}{N\Delta t}\right)$:

$$F(\frac{n}{N\Delta t}) = \sum_{k=0}^{N-1} f(k\Delta t) e^{-i2\pi kn/N} \tag{5}$$

where N is the number of sampling in a cycle. Δt is time interval.
Then, follows the wave magnitude $\left|F(\frac{n}{N\Delta t})\right|$ and phase φ, we have

$$F(\frac{n}{N\Delta t}) = \left|F(\frac{n}{N\Delta t})\right| e^{i\varphi(\frac{n}{N\Delta t})} = F_n \tag{6}$$

Inverse the frequency domain F_n to time domain function $f(k\Delta t)$. By summing up the expanded series, the travel time value is approximately estimated (Kreyszig) [8]:

$$f(k\Delta t) = \frac{1}{N}\sum_{n=0}^{N-1} F_n e^{i2\pi kn/N} \tag{7}$$

How to determine that N is accepted? This can be verified by the Chi-square test upon the given level of significance, α.

(3) Kalman filter:
Kalman filter is a linear filter of sequentially recursive equations. Given that $\theta(k)$ is the system state variable at the k^{th} step, $x(k)$ is the state coefficient, $n(k)$ is the coefficient of the system noise $w(k)$. $\delta_1(k)$ is the covariance matrix of $w(k)$. $y(k)$ is the system measurement/ observed value. $m(k)$ is the coefficient of the measurement. $\delta_2(k)$ is the covariance matrix of the measurement error term $v(k)$.

The system equation

$$\theta(k+1) = x(k)\theta(k) + n(k)w(k) \tag{8}$$

The measurement equation

$$y(k) = m(k)\theta(k) + v(k) \tag{9}$$

Given an initial value $\theta(0)$ and error variance matrix $P(0)$, by means of updating the observed data one by one, $y(k), (k = 0, 1, 2\ldots)$, cooperating with the filtered value at the last state $\theta(k-1)$, and error variance matrix $P(k-1)$, then the filter value at current state $\hat{x}(k)$ and $P(k)$ are obtained. The iterative relationship reveals as follows:

$$\hat{\theta}(k|k) = \hat{\theta}(k|k-1) + K(k)\left[y(k) - m(k)\hat{\theta}(k|k-1)\right] \tag{10}$$

$$\hat{\theta}(k+1|k) = x(k)\hat{\theta}(k|k) \tag{11}$$

$$K(k) = p(k|k-1)m^T(k)\left[\delta_2(k) + m(k)P(k|k-1)m^T(k)\right]^{-1} \tag{12}$$

$$P(k+1|k) = x(k)P(k|k)x^T(k) + n(k)\delta_1(k)n^T(k) \tag{13}$$

$$P(k|k) = [1 - K(k)m(k)]P(k|k-1) \tag{14}$$

in which, $\hat{\theta}(k+1)$ expresses the forwarding prediction of $\theta(k)$. $\hat{\theta}(k|k)$ denotes the filter value or estimator of $\theta(k)$. $K(k)$ is the gain matrix.

As previous, Kalman filter has two stages for processing. One is to get filter value, and another is to make a prediction.

(4) α-β-γ filter:
Kalman filter in fact has its difficult initially on setting up a correct system state. Instead, this paper introduces a second order difference equation, said α-β-γ filter, to track the system state of the trajectory. However, performing prediction, two

primary variance matrices should be initiated for the system noise and measurement error.

Assuming that the measurement/observed data y, sequentially as time t in discrete interval expression of k: $y(0), y(1), \ldots y(k-1)$, the measured/observed value is:

$$y(1) = s(k) + v(k) \tag{15}$$

where, $s(k)$ is a uniformly accelerative motion (constant), and $v(k)$ is the noise with variance $E[v^2(k)] = r_2$, a zero mean white noise series. The second order difference is constant.

Herein, how to estimate $s(k)$ through the data $y(0), y(1), \ldots y(k)$, is described. Supposing that t is sufficient small, by the Taylor's Expansion method, $s(k)$ is approximate as:

$$s(k) = s(k-1) + ts(n-1) + \frac{t^2}{2}s(k-1) \tag{16}$$

For simplicity, Eq. (16) can be rewritten as:

$$\theta(k) = x\theta(k-1) + n\delta_1(n-1) \tag{17}$$

$$\theta(k) = \begin{bmatrix} s(k) \\ s'(k) \\ s''(k) \end{bmatrix}, x = \begin{bmatrix} 1 & t & \frac{t^2}{2} \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix}, n = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$\delta_1(k-1)$ has a variance $E\left[\delta_1^2(k)\right] = r_1$ with zero mean white noise series. Equation (17) obviously specifies,

$$\theta(k) = \begin{bmatrix} s(k) \\ s'(k) \\ s''(k) \end{bmatrix} = \begin{bmatrix} 1 & t & t^2/2 \\ 0 & 1 & t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s(k-1) \\ s'(k-1) \\ s''(k-1) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \delta_1(k-1) \tag{18}$$

Equation (18) means a time invariant system. Then, via Kalman filter manipulation, the next state can be inferenced recursively. Indeed, this way can be improved up to the third order, the forth order…

2.2 Travel Time Prediction

Along with applications, travel time prediction can be classified into forward forecasting (FF) and backward prediction (BP). FF is knowing the starting time and to forecast the coming/future states. BP is the ending time be given and backward estimate what the available starting time be set.

A. Forward forecasting:
   Firstly input the original, destination, and the scheduled clock time to start the trip. The computation is to, sequentially sum up the estimated travel time in the

term of the arrival clock time on each link of the route from the original to the destination. Referring to [9], if the cumulative sum is below 15 min, Kalman filter approach works for the estimation. If the sum accumulates over 15 min, a long term travel time prediction model is executed. As previous, Fourier Transform applies to historical data to extract out the reliable parameters for long term prediction. The calculation is executed till reaching to the destination.

B. Backward prediction:
Firstly input the original, destination and expected arrival time. The computation is to sequentially sum up the estimated travel time in the term of the corresponding clock time traveling on the links from the destination backward to the original. All the calculation depends on historical data, operated with the Fourier Transform method. The result is to come out a suggested starting time for travel.

3 Experiment and Discussion

The test data set is got from Taiwan Freeway, particularly on the northern segment and the Taipei airport express connector. The data records each with 15 min for an interval. There are 38 sampling days for the following experiments.

3.1 Data Compensation Experiments and Analysis

Because holiday and workday have different characteristic, this study divides them in separated group for analysis. There are four experiments respectively for pretime compensation and realtime compensation. Each experiment has set a random loss of 5%, 10%, 15%, … respectively. Every loss will be simulated of 30 times. Pretime compensation has three methods: $F_1(t)$ = cubic spline curve interpolation, $F_2(t)$ = linear interpolation, $F_3(t)$ = near value method. Realtime compensation has four methods: $R_1(t)$ = average value method, $R_2(t)$ = Fourier transform method, $R_3(t)$ = α-β-γ filter cooperated with average value method, $R_4(t)$ = α-β-γ filter cooperated with Fourier transform method. Finally, we take the MAPE (mean average percentage error) analysis for every experiments to explore their efficiency. Figures 1 and 2 illustrate the results.

According to the experiments, we find that about the pretime interpolation has:

(1) On workdays, $F_1(t)$ and $F_2(t)$ are better than $F_3(t)$, and their MAPEs are all lower than 10% while the loss rate below 60%. This concludes that they are belonging to high precision methods.
(2) On holidays, $F_1(t)$ and $F_2(t)$ are better than $F_3(t)$ while the loss rate below 60%.
(3) The MAPEs of $F_1(t)$ and $F_2(t)$ are smaller than 10% while their loss rate smaller than 90%.
(4) According to the data base, the testing road segments' loss rate, in fact is averagely at 16.2% on workdays, 12.6% on holidays, respectively. However, in the tests about pretime interpolation, the MAPE of $F_2(t)$ on workdays is averagely 7.38% while its loss rate is lower than 20%, and on holidays is averagely 4.36%. This leads to that using linear interpolation can achieve the level of high precision.

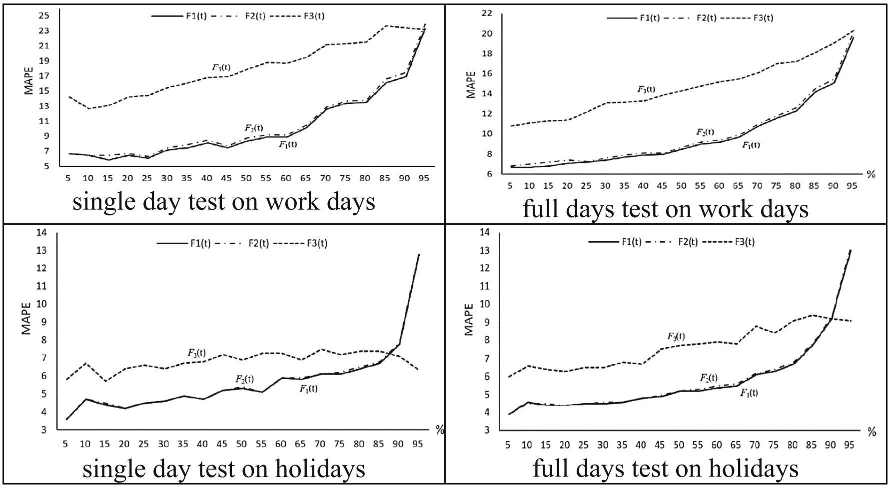


Fig. 1. MAPE value of pretime interpolation

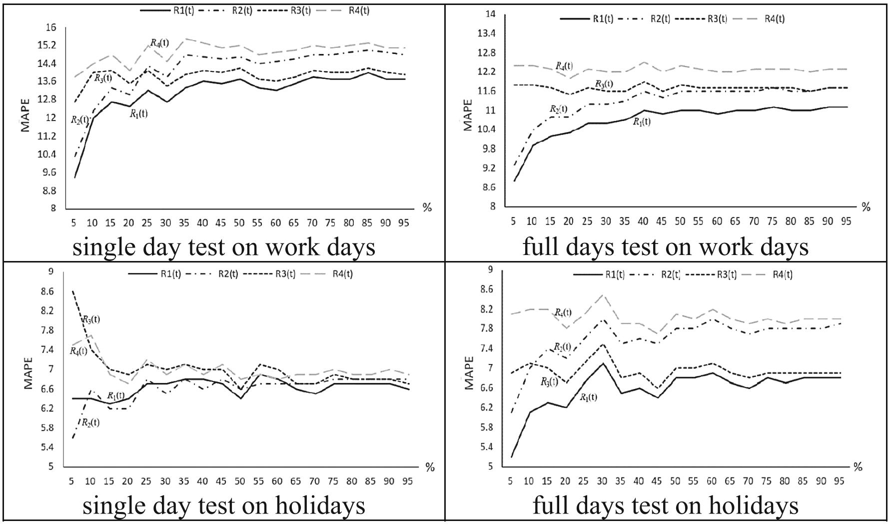


Fig. 2. MAPE value of realtime extrapolation

As for the realtime compensation:

(1) On workdays, the efficiency of $R_1(t)$, $R_2(t)$ and $R_3(t)$ almost has equivalent quality.
(2) On holidays, there are no evident rush hours, then $R_1(t)$, $R_2(t)$, $R_3(t)$, $R_4(t)$ present the same efficiency.

(3) In comparison with $R_2(t)$ and $R_1(t)$, $R_2(t)$ has less storage space than $R_1(t)$. When processing hardware needs exchanging large data, $R_2(t)$ (Fourier transform method) will be saver than $R_1(t)$ (average value method) in communication packet load. This concludes that $R_2(t)$ can save money expenditure.
(4) In comparison with $R_3(t)$ and $R_4(t)$, $R_4(t)$ has the advantage of $R_2(t)$ and it working with α-β-γ filter presents out the compensation just in time.
(5) According to the data base, the testing road segments' loss rate, in fact is averagely at 16.2% on workdays, 12.6% on holidays, respectively. However, in the tests about realtime compensation, the MAPE of $R_1(t)$ on workdays is averagely 10.26% while its loss rate is lower than 20%, and on holidays is averagely 6.20%. This leads to that using realtime compensation can tends to the level of high precision.

3.2 Travel Time Prediction Experiment and Analysis

This study uses probe vehicles to get field test samples. About short range of travel time prediction (define the range within 30 min), two tasks are done: applying pretime interpolation to finish a complete historical data base and adopting realtime extrapolation to make the prediction. Four routes for the practices are involved, each route randomly takes 30 probe vehicles for comparison. All results of MAPE lay within 10%, revealing high precision of the propositions.

About long range prediction, the result of MAPE is about 8%, revealing high precision.

4 Summary/Conclusion

This paper organizes and builds up the data compensation mechanism. In which, pretime interpolation is for daily maintenance to finish completeness of historical data. Realtime extrapolation is for online travel time prediction.

Based on the probe vehicles' returning data, it proves that the cubic spline curve and linear interpolation methods are acceptable. In considering the number of data and computing efficiency, the study suggests the linear method.

With respect to realtime extrapolation, this study found no exactly difference among four methods, no matter when they for workdays or holidays. About travel time prediction, no matter what methods for short time or long time prediction, all can work well and acceptable. And, they all present robustness.

References

1. Lin, D.J., Ai, J.M., Chen, Z.X., Zheng, J.H., Wu, Y.Z.: Traffic data stitching based on ARIMA model. Transp. Inf. Saf. (in Chinese) **33**(3), 77–81 (2015)
2. Guo, M., Lan, J.H., Li, J.J., Lin, Z.S., Sun, X.R.: Traffic flow data recovery algorithm based on gray residual GM (1, N) Model. J. Transp. Syst. Eng. Inf. Technol. (in Chinese) **12**(1), 42–47 (2012)
3. Wang, X.X., Fu, H., Yang, M.J., Lu, J.F.: Study on the method of completing missing dynamic traffic data. J. Inner Mongolia Agric. Univ. (in Chinese) **37**(3), 93–96 (2016)
4. Yin, L.S., He, Y.G., Dong, X.P., Lu, Z.Q.: Research on the multi-step prediction of volterra neural network for traffic flow. Acta Automatica Sinica **40**(9), 2066–2072 (2014)
5. Peeta, S., Anastassopoulos, L.: Automatic real-time detection and correction of erroneous detector data with Fourier transforms for online traffic control architectures. Transp. Res. Rec. **1811**, 1–11 (2002)
6. Chang, T.H., Huang, H.R.: On-line reparation for data loss of vehicle detectors. J. Transp. (in Chinese) **20**(4), 300–320 (2008)
7. Barry, P.J., Goldman, R.N.: A recursive evaluation algorithm for a class of Catmull–Rom splines. In: Proceedings of the 15st Annual Conference on Computer Graphics and Interactive Techniques, pp. 199–204, August 1988
8. Kreyszig, E.: Advanced Engineering Mathematics. Wiley, Hoboken (1999)
9. Zhang, T.X., Guo, Z.T.: A prediction model for bus arrival time and multi-route travel time. China Civil Eng. J. **38**(12), 115–123 (2005)

Wheelchair Control System Based on Electromyography

Zhi-Jie Hu(✉), Hao Li, Xiu-Fang Zhang, Chang-Hao Li, Xin Wang, and Miao-Na Zhang

Xi'an Eurasia University, Xi'an, ShaanXi, China
{huzhijie, lihao, zhangxiufang}@eurasia.edu

Abstract. In order to make life more pleasant and convenient for people with physical disability and mobility problems, this paper designs a wheelchair operation system based on electromyogram. This design extracts four different gestures of the human body to control the four movements of the wheelchair, namely forward, backward, left and right. In order to ensure the user's safety, this design uses ultrasonic distance sensor to detect the road condition, avoid some obstacles in time, and make the wheelchair move forward safely. This design mainly includes EMG signal collection, analysis, classification and identification, wireless transmission, EMG signal processing, wheelchair control system, safety control system, etc. This design has the advantages of simple operation, high security, stable signal and strong anti-interference, etc.

Keywords: EMG signal · Feature extraction · Classification identification
Wireless transmission · Safety control

1 Introduction

Wheelchair is an important tool for rehabilitation. It is not only a walking tool for people with physical disability and mobility problems, but also an important tool for them to physical exercise and participate in social activities.

China's population aging and consumption upgrading and other endogenous factors drive the rapid growth of wheelchair industry. The aging of the population will exist for a long time, which will promote the demand for wheelchairs, and the consumption upgrading will promote the development of high-end wheelchair market.

With the aggravation of China's aging population and the improvement of people's living standards. The explosive power of the elderly market has been prominent in recent years. According to China wheelchair industry survey and analysis and development trend forecast report (2017–2023) released by China industry research network, China's wheelchair industry has maintained rapid development in recent years. However, the product structure of the wheelchair industry is single, most products for the general wheelchair, the product price is low. High-end products such as electric wheelchairs and sports wheelchairs are underinvested in research and development, and their technical level is quite different from that of foreign countries. Sports wheelchairs are mainly imported.

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 337–342, 2019.
https://doi.org/10.1007/978-3-030-04585-2_41

Nowadays, surface electromyography signals have been widely used in the equipment of biomedical robots, including the diagnosis of neurological diseases and the control of medical auxiliary equipment. Compared with other bioelectric signals, surface electromyography signals are easy to collect and have large signal amplitudes. Conventional signal collection method is that many complex electrode wires connect the user and collection devices, as a result, the user and collection equipment are confined together in the application, resulting in electromyography greatly reduces the scope of application. Based on the above problems, this design proposes a wireless wheelchair control system based on electromyography signal.

EMG-controlled wheelchair is a trend from Internet to Internet of things (IoT) under the influence of Internet. EMG-controled wheelchair are compared with ordinary wheelchair, it not only keep the traditional wheelchair electric function, but also realize the wheelchair automation control by another way—EMG-controled mode, solved people with disabilities control wheelchair, only the hand action of change can control the movement of the wheelchair. Control system of EMG-controled wheelchair will become a new type of control mode with high comfort, high security and high convenience.

2 Design of System

Firstly, in this design, a surface electromyography (sEMG) acquisition platform was built, including the design of the experimental process and the construction of the experimental environment. Secondly, the command of gesture control is extracted from the sEMG of human body, including the preliminary filtering analysis, feature extraction and action classification of EMG. Finally, a wheelchair control system based on EMG is designed to realize the purpose of controlling the wheelchair directly by human actions.

This design system block diagram is shown in Fig. 1, mainly including the EMG collection module, EMG analysis module, EMG recognition module, wireless transmission module, signal processing module, wheelchair control module and safety control module. EMG acquisition module will collect on the human body EMG and send them to EMG analysis module to analyze, through preamplifier, photoelectric isolation, high-pass filtering, main amplification, low-pass filtering and notch filtering processing after converted into digital signal by A/D module. The EMG analysis module extracts a set of data features that can represent the inherent characteristics of gestures from the EMG and input relevant data into the EMG sensor. When the user makes a gesture, the sensor can recognize the relevant electromyographic signal, return dynamic or static information to the microprocessor, and other irrelevant gestures cannot be recognized. The information is transmitted to the microprocessor via the wireless transmission module for data processing, while the signal processing module also processes the data from the security control module. The processed data is transferred to the wheelchair control module, and the wheelchair control module receives data instructions. According to these instructions, the wheelchair is controlled to make forward, backward, steering, emergency stop and other actions.

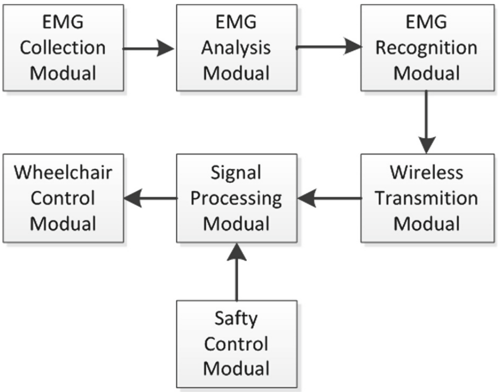


Fig. 1. System block diagram

3 Design Content

3.1 EMG Collection Module

According to difference of the guide electrode used in the measurement and the placement, EMG collection can be divided into Needle EMG and Surface EMG. The former is a needle electrode and the latter is a surface electrode. Surface electrode was used to collect surface EMG in this design. The surface EMG is the result of complex subepidermal electromyography in the time and space of skin surface during muscle contraction.

3.2 EMG Analysis Module

A. Gesture recognition
   The EMG analysis module mainly uses EMG signal for gesture recognition. At present, there are four embedded gestures: clench, extend palm, flex wrist and extend wrist. These four gestures correspond to four wheelchair driving directions: forward, backward, left and right. The corresponding gesture recognition table is shown in Table 1.

Table 1. The corresponding gesture recognition table

Byte [0]	Gesture	Wheelchair directions
0x00	Relax	Static
0x01	Clench	Forward
0x02	Extend palm	Backward
0x03	Flex wrist	Left
0x04	Extend wrist	Right

B. Signal condition
SEMG is a weak non-stationary random signal, which is easy to be interfered by external noise sources and has low signal-to-noise ratio. In order to increase the signal-to-noise ratio, multichannel SEMG signal must be amplified and filtered by signal conditioning circuit. In general, signal is detected by the surface electrodes that needs to pass through preamplifier, photoelectric isolation, high-pass filtering, main amplification, low-pass filtering and notch filtering processing after converted to digital signal by A/D module. The amplification factor of preamplifier is only ten to dozens times, while the signal of preamplifier is amplified to volt level by main amplification. High - pass filtering is mainly used for isolating and filtering low - frequency interference (baseline drift). Isolation is to isolate the human body from electrical equipment to ensure the safety. Low - pass filters are used to pass low - frequency signals, attenuate or suppress high - frequency signals. Notch filters are generally used to filter out 50 Hz power frequency interference.

C. Signal extraction
The main target of EMG extraction is to extract a set of data characteristics that describe intrinsic properties of this gesture. As far as possible, the features extracted from different actions have obvious differences in order to facilitate classification. This design adopts the wavelet analysis method to carry on EMG feature extraction. The wavelet analysis method is a combination of time domain and frequency domain analysis method, it is the new development of Fourier transform and has variable analysis window the time domain and frequency domain. It acts like a bandpass filter with constant bandwidth and variable center frequency, uses short window in high frequency and use the wide window in low frequency. This provides a reliable way for real-time signal processing.

3.3 Wireless Transmission Module

Considering that bluetooth has the advantage of low power in short distance (maximum transmission distance is 10 m), this design uses bluetooth for wireless transmission.

Bluetooth transmission module receives electrical signals through the EMG signal receiver receiving neuromuscular system in the brain that controls the hand make different moves, after through processing module, the oscillator module. According to oscillator module, EMG signals are transmitted to the bluetooth receiver installed on intelligent wheelchair. Then bluetooth receiver transmits the processed signal to the master control module, completing a wireless transmission, and then circulates the whole process again when the next EMG signal arrives.

3.4 Signal Processing Module

A. Active segment detection of signal
The purpose of detecting the active segment of EMG signal is to extract the signal corresponding to the action execution from the data stream of continuous multichannel EMG signal. The task of active segment detection is to determine the starting and ending positions of SEMG signals caused by muscle movement.

Accurate identification of SEMG signal active segment detection is a prerequisite for feature extraction and classification. At present, the most commonly used SEMG active segment extraction algorithms include short time fourier transform (STFT), absolute value MAV feature, moving Average Method, etc.

B. Short time fourier transform

In this design, short - time Fourier transform is used to detect signal active segment. Short - time Fourier transform can achieve frequency localization in certain degree. Short time Fourier transform is through the window of time g(t − r) to get the time r signal characteristics, it is the average of time signal characteristics in time widow, the shorter the signal in the time window, the higher the time resolution. The expression is as follows:

$$STFTF(\omega, \tau) = \int R\mathrm{f}(t)g * (m - \tau)e - jwtdt \tag{1}$$

The idea is: choose a window function of time-frequency localization, assume analysis function g(t) is smooth in a short time interval, move window function f(t) g(t) is stationary signal in different limited width, so as to calculate the power spectrum of each different time. The short time fourier transform uses a fixed window function. Once the window function is determined, its shape will not change, and the resolution of the short time fourier transform will be determined. If you want to change the resolution, you need to reselect the window function. The short time fourier transform can be used to analyze stationary signal or approximate stationary signal, but for non-stationary signal, when the signal transformation is intense, the window function is required to have a high time resolution. When the waveform transformation is relatively flat, the low-frequency signal mainly requires the window function to have a high frequency resolution.

3.5 Wheelchair Control Module

After EMG signal collecting sensors complete the collecting signals, the EMG signals are transmited to the microprocessor controller via Bluetooth. Then microprocessor controller can process signals and convert the signal into a signal that the wheelchair control module can recognize and manipulate the wheelchair.

This design chooses LGT 8F328P as micro controller, the controller has a high performance and low power eight bit LGT8XM kernel, senior RSIC architecture, 32 kbytes on online programming FLASH program memory, 2 kbytes internal data SRAM, working voltage from 1.8 V to 5.5 V.

3.6 Safety Control Module

After EMG signal collecting sensors complete the collecting signals, the EMG signals are transmited to the microprocessor controller via Bluetooth. Then microprocessor controller can process signals and convert the signal into a signal that the wheelchair control module can recognize and manipulate the wheelchair.

Installed in wheelchair front-end ultrasonic distance sensor can determine whether the distance between itself and the obstacles ahead is less than the safe distance, and then sends a signal to main control board to control the wheelchair to stop forward, in order to ensure security when the user can't control the wheelchair under emergency. Wheelchair rear ultrasonic distance sensor can be used to remind and warn the user from the rear of the unknown security hidden danger, if less than the safe distance of setting from the rear of the obstacles, the rear of the ultrasonic distance sensor sends a signal to main control board to control the wheelchair for safety to avoid, so as to ensure the safety of users.

4 Conclusions

The wheelchair control system based on EMG use arm muscles of users to transmit electrical signals for controlling the wheelchair. Compared with the traditional way of wheelchair control, the convenience of wheelchair control is improved and the safety of users is guaranteed. It has very important practical significance and market value.

Acknowledgement. The authors are grateful to anonymous reviewers for their comments, which may help us to greatly improve the article. This work was supported by Science category of university-level projects in 2018 of Xi'an Eurasia University (Grant numbers: 2018XJZK11).

References

1. Yao, P.F., Sheng, X.J., Guo, W.C., Hua, L., Zhu, X.Y.: A humanoid prosthetic hand control system based on the co-decoding of surface electromyography and near infrared spectroscopy. Chin. J. Rehabil. Med. **31**(5), 547–551 (2016). (in Chinese)
2. Li, F., Chen, B.: Design and application of wearable motion analysis system based on electromyography. Consum. Electron. 43–46 (2017)
3. Zhao, Z.Y.: Study on several key techniques in the detection and processing of surface electromyography signal. University of Science and Technology of China (2010)
4. Zou, X.Y., Tan, B., He, Q.Y.: Low power consumption of BLE bluetooth in wireless sensor nodes. Mod. Electron. Technol. **39**(20), 67–71 (2016). (in Chinese)
5. Mao, D.J.: Chip system design of surface electromyographic signal collection and transmission based on bluetooth low power consumption. Zhejiang University (2017)

Using On-Board Diagnostics Data to Analyze Driving Behavior and Fuel Consumption

Chao-Fu Yeh(✉), Liang-Tay Lin, Pei-Ju Wu, and Chi-Chang Huang

Feng-Chia University, No. 100, Wenhua Rd. Xitun dist., Taichung, Taiwan
cfyeh@fcu.edu.tw

Abstract. The reduction of carbon use on the road traffic seems obviously to be a right direction and important strategy in the worldwide environment. In the road traffic, the bus operation belongs to a commercial vehicle related to longer travel time and distance, thus, it deserves to pay more attention on the fuel consumption of bus operation in order to reduce the air pollution emission and increase the energy-saving efficiency. Although bus transport systems contain huge operations data, there is little practical knowledge of how to make use of the data. Hence, this study aims to explore big data of bus transport systems and create valuable environmental operations strategies.

Our research aims at using the second generations of on-board diagnostics system (OBD II) to output the real and dynamic data of engine oil consumption. Our research focus on the studying on the eco-driving behavior of bus based on the data from OBD II in order to improve the management of energy-saving for bus operators. In final, the results of research show that there is positive correlation in statistics between the speed, the engine temperature, the ambient air temperature, travel distance and energy consumption. These five variables are associated with the oil consumption.

Keywords: Energy saving · OBD II · Driving behavior · Eco-driving analysis

1 Introduction

Reducing fuel consumption and greenhouse gas (GHG) emissions is one major approach towards controlling global warming. In Taiwan, the greenhouse gas emissions (GHG) in various sectors in 2013, the transportation sector is secondary for CO_2 emission, especially for the road traffic (approximately 95% of the total energy consumption on transportation sector). In the road traffic, the bus operation belongs to a commercial vehicle related to longer travel time and distance, thus, it deserves to pay more attention on the fuel consumption of bus operation in order to reduce the air pollution emission and increase the energy-saving efficiency.

Evaluating driver performance and promoting energy efficient driving has received scarce attention from the research community. This is due to the difficulty of objectively evaluating human driver performance. The driver controls the speed, acceleration, braking, engine rotation speed, the gear engaged (Reiter 1991), and the position of the vehicle on the street in an environment characterized by certain traffic conditions, itinerary, load, etc. Different driving styles result in different fuel consumption levels,

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 343–351, 2019.
https://doi.org/10.1007/978-3-030-04585-2_42

thus related to driving efficiency. Different external conditions result in different levels of consumption.

Our research aims at using the second generations of on-board diagnostics II system (OBD II) to output the real and dynamic data of engine oil consumption. Then, it could induce the eco-driving behavior based on the instantaneous data of engine oil consumption. Eco-driving aims to change the driving behavior through simple advice such as maintain a steady speed, accelerate moderately, follow speed limits, anticipate traffic flow and maintain your vehicle (Stæhr 2013).

Past research neglected the factors influencing bus energy consumption, pollution generation, driving behaviors and bus lines energy strategies. With the growing popularity and maturity of On Board Diagnostics (OBD), more attention has been directed to using existing vehicle mounted equipment for data capture and analysis to study improvements in energy efficiency. Thus, It is a basic assumption that the On-Board Diagnostics data provides ground-truth results, meaning that the recorded values are assumed to be correct. The main contributions of this paper are: (1) First experimental evaluation of fuel efficiency based on a large set of OBD II data; (2) Analysis the relationship between the parameters of OBD and driver's behavior; (3) Concrete advice on how to further improve the management of eco-driving.

2 Literature Reviews

This research aims at analyzing the relationship between the fuel consumption and the parameters of driving based on the data from on-board diagnostics. From the review of literatures, we find that researches had done many efforts to explore the relationship between the influence factors of driving behavior and eco-driving.

Evans (1979) indicated that if a driver conducted less acceleration and mildly drives a vehicle – and if the time for a trip remains the same – 14% of fuel consumption can be saved. Waters and Laker (1980) used a real vehicle to experiment on the influences of a driver's habits on fuel consumption and the result showed that mild driving behaviors could reduce up to 15% of fuel consumption. Hooker (1988) investigated factors influencing fuel consumption and tried to construct the most economic driving patterns. The result showed that gear change timing, driving speeds, urgent acceleration were significant influential factors.

Ericsson (2000 and 2001) studied fuel consumption and pollution emissions and chose to focus on the correlations between driving patterns and fuel consumption. He found that a vehicle running at 50–70 km/h without urgent acceleration or braking would significantly reduce fuel consumption.

Alessandrini (2006) mentioned that use of vehicles could induce power requirement, fuel consumption and Cox emissions. He compared those data with the vehicle chassis dynamometer testing method and found a less than 3% difference between the data retrieved by OBD and OBD-II, which confirmed that OBD-II is a viable method in studies on fuel consumption.

Berry (2007) indicated that a study on 10 vehicles over 10 months for fuel consumption at high speeds and at aggressive accelerations. The result of his research mentioned reduction in fuel consumption of 5.8%. Mentioned that monitoring the drivers' fuel consumption after 12 months of eco-driving training. The result of observation was average reduction in fuel consumption of 2% after eco-driving training.

Browne (2008) used OBD-II to capture some parameter values on driving conditions and hoped to direct the attention of a driver to variables that could influence fuel efficiency such as the type of air cleaner, the driving speed, the rotational speeds of the engine and the load capacity. He derived from the ANOVA table that the driving speed, the rotational speeds of the engine and the type of air cleaner were all factors significantly correlated with driving behaviors.

Beusen (2009) observed the driving behavior of 10 drivers and analyzed the fuel consumption based on the real data during 10 months. The research findings were the reduction in fuel consumption of 5.8% for first 4 months. Analyzed the data offered from by bus operators in Taiwan and indicated that drivers who participated in eco-driving courses could change their habits and save up to 10% to 15% of energy consumption.

Used OBD-II to build in various devices for surveillance of the vehicle driving conditions to collect parameter values and build fuel consumption models. OBD-II can monitor and control more than 80 types of driving conditions such as the driving speed, the rotational speeds of the engine, the temperature of the cooling water, and the battery voltage.

Attempted to describe Driving Cycle Performances via OBD-II to construct fuel consumption patterns. His study showed that as rainy days lowered the friction force on the ground, vehicle acceleration would only require more fuel consumption but exhibit less efficacy. Considered idling and other circumstances that caused a vehicle to idle or stop as the most significant challenges to fuel-saving goals. Monitored the driving behaviors of 20 drivers for saving the fuel consumption and obtained the result for reduction in fuel consumption of 4.8%.

3 Methodology

To better understand bus driving behavior, this study adopted vehicle GPS information and parameter values from OBD-II (On-board Diagnostics System II) to conduct relational analysis. Moreover, this study investigated the relationships among drivers, vehicles and roads in order to analyze the relationship between driving behaviors and fuel consumption, such as rotational speeds of engine, idling, temperature of engine, and voltage. Furthermore, GPS information and vehicle conditions can be utilized to identify appropriate routes. As shown in Table 1, the data could be obtained from OBD-II, e-Bus, E-Ticket System and Geographic Information System (GIS) data. The data format is also shown in Table 1.

One of the most efficient approaches to evaluate driver performance is to register a set of events (parameters) read from the on-board diagnostics, which stores messages from all driving events on on-board recorder, from where data is retrieved and stored in

Table 1. Data format and sources

No.	Data facilities (sources)	Data formats and available parameters
1	e-Bus output data	Date in use, bus operators, routes (route codes), drivers, vehicle numbers, departure time, returning time, time for arrival and departure at each stop, paths
2	Ticket information output data	Date in use, routes, departure stop, destination stop, vehicle numbers, ticket company, ticket type, card number, boarding time, boarding bus stop, alighting time, and alighting bus stop
3	On board diagnostics (OBD II)	Computer time, engine temperatures, engine load, engine torque, rotational speeds (per minute), speed, trip distance, fuel oil consumption, instantaneous oil consumption, and ambient temperature
4	Geographic information system (GIS)	Roads, railways, MRTs, administrative areas, rivers, lakes, and landmarks over that map layers (institute of transportation, ministry of transportation and communications), paths(lengths, routes), and slope (Google map)

a database for subsequent analysis. On-board diagnostics (OBD) is a device monitoring the pollution emission and the efficacy of engine in the vehicle. Besides, it can obtain so rich data related to the speed, rotational rate, fuel consumption and distance etc. through OBD device. The work methodology of the research presented in this paper involves three main processes as suggested in Fig. 1.

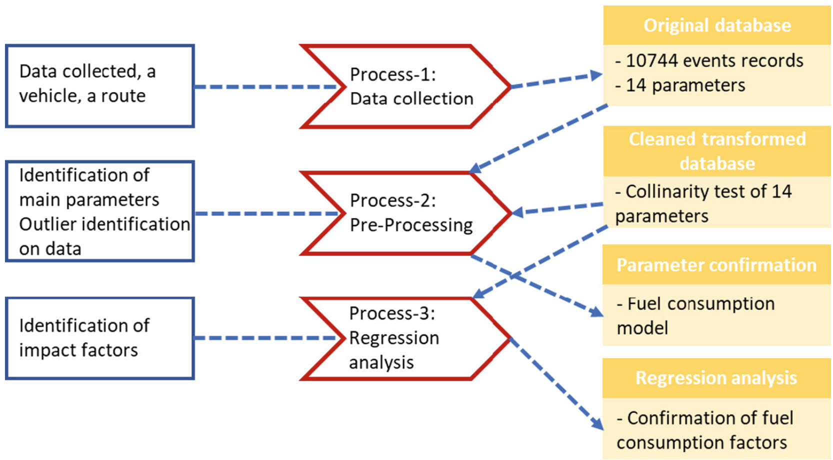


Fig. 1. Work methodology of research

4 Analysis Results

The research can be divided into the data preprocessing and data analysis. First the data preprocessing aims at integrating diverse data and verifying the independence of variables in order to understand well the data structure. Through factor analysis and multiple regression analysis, it can find the key factors of fuel consumption. Then, using the regression analysis in SPSS displays the analysis results.

4.1 Data Preprocessing

Raw data includes 53715 data from on-board diagnostic system in vehicle in order to obtain the data of fuel consumption per 5 s. To avoid much fragmental data, it has to integrate and average raw data by every 5 s. In addition, it can take 14 different parameters from OBD II. Our research chooses 5 important parameters to analyze the relationship between the driving behaviors and fuel consumption in order to explore key factors of eco-driving. The definition of relative parameters as following:

1. Engine torsion: engine thrust (unit: Kg/m)
2. Rotation rate per minute: average rotation rate per minute (unit: RPM)
3. Speed: average speed (unit: km/h)
4. Temperature of engine: Temperature after engine working (unit: °C)
5. Trip distance: average distance of trip (unit: km)

4.2 Data Analysis

First, it has to describe the definition of variables and construct forecasting model of fuel consumption based on the regression analysis in SPSS. Our research divides the data from OBD II into three categories of independent variables: driver operation, external environment and vehicles. The attribution and definition of independent/dependent variables are listed in the Table 2.

Table 2. Definition of driving variables

Attribution	Variable	Typology	Code	Definition	Unit
Driving operation	Engine torsion	Continual variable	x1	engine thrust	Kg/m
	Rotation	Continual variable	x2	average rotation rate per minute	RPM
	Speed	Continual variable	x3	Average speed	km/h
Environment	Engine temperature	Continual variable	x4	Temperature after engine working	°C
	Around air temperature	Continual variable	x5	Air temperature around the vehicle	°C
Vehicle	Trip distance	Continual variable	x6	Average distance	km
Fuel consumption	Fuel consumption rate	Continual variable	x7	Fuel consumption per kilometer	km/L
	Fuel economics	Continual variable	x8	distance per liter of fuel	km/L
	Average fuel economics	Continual variable	x9	distance from starting engine to closing engine divided by total fuel consumption	km/L

Our research aims at using multi-regression model to analyze the relationship between driving operation and fuel consumption. The regression analysis of engine torsion and fuel consumption shows as Eq. 1.

$$\hat{y} = \beta_0 + (\beta_1 x_1) \tag{1}$$

In which, β_0 is constant; β_1 is coefficient of variable X_1.

Using the multi-regression analysis of virtual variable builds up the forecasting model of fuel consumption, then it can try to find the solution by regression step by step. The result of analysis of variance (ANOVA) is shown in the Table 3. According the F value, the result is significant. It means that this regression analysis (engine torsion) can efficiently explain the forecasting of fuel consumption.

Table 3. ANOVA analysis of engine torsion (linear regression)

Model	Sum of squares	df	Mean square	F	Sig.
Regression variance	2270.283	4	2270.283	1322.728	0.000
Residual variance	18437.190	10742	1.716		

About the fuel consumption model, the coefficient of analysis has presented as Table 4. P-value of significant variance is less significant level $\alpha = 0.05$, it can confirm the forecast model of fuel consumption, as Eq. 2, in which there is positive correlation between engine torsion and fuel consumption.

$$\hat{y} = 4.942 + 0.024x_1 \tag{2}$$

Our research also aims at using linear-regression model to analyze the relationship between environment and fuel consumption. The regression analysis of engine torsion and fuel consumption shows as Eq. 3.

$$\hat{y} = \beta_0 + (\beta_1 x_4) \tag{3}$$

In which, β_0 is constant, β_4 is the coefficient of variable X_4.

Table 4. Coefficient model of engine torsion (linear regression)

Model	Unstandardized coefficients		Standardized coefficient	t	Sig.	Collinearity statistics (tolerance)
	B	Std. error	Beta			
Constant	4.4942	0.019		260.599	0.000	
Engine torsion	0.024	0.001	0.331	36.369	0.000	1.000

According to F value, the statistical result is significant. It means that the variable "Engine temperature" can efficiently explain the forecasting of fuel consumption. Concerned the fuel consumption model, P-value of significant variance is less the level of significant $\alpha = 0.05$, the analysis result can be presented the forecast model of fuel consumption, as Eq. 4, in which it shows positive correlation between engine temperature and fuel consumption.

$$\hat{y} = 12.613 + 0.225x_1 \tag{4}$$

Our research aims at using multi-regression model to analyze the relationship between vehicle and fuel consumption. The regression analysis of engine torsion and fuel consumption shows as Eq. 5.

$$\hat{y} = \beta_0 + (\beta_1 x_5) \tag{5}$$

In which, β_0 is constant, β_5 is the coefficient of variable X_5.

The statistical result is significant based on F value is 432227.651. It means that the variable "Trip distance" can efficiently explain the forecasting of fuel consumption. Concerning the fuel consumption model, P-value of significant variance is less the level of significance $\alpha = 0.05$, the analysis result can be presented the forecast model of fuel consumption, as Eq. 6, in which it shows positive correlation between trip distance and fuel consumption.

$$\hat{y} = -2476.278 + 0.038x_1 \tag{6}$$

The objective of research aims at using the regression analysis to build up the relationship on the fuel consumption and driving variables in order to confirm the influence factors of eco-driving. Thus, the research first divides the variables into three attributions (such as driving operation. Environment and vehicle) and approves the correlation between the variables and fuel consumption. The result of regression analysis is summarized in Table 5.

Table 5. Summary of regression analysis on fuel consumption

Attribution	Variables	Correlation	Influence factor
Driving operation	Engine torsion	Positive	0.024
	Rotation	Positive	0.002
	Speed	Positive	0.028
Environment	Engine temperature	Positive	0.225
	Around air temperature	Negative	0.173
Vehicle	Trip distance	Positive	0.038

Following the statistical analysis above, there are 5 potential variables which can affect the fuel consumption, such as engine torsion, rotation, speed, engine temperature and trip distance. These 5 variables can also link to the driving behavior, the description of relationship between the variables and driving behavior is as follows.

1. Engine torsion: meant engine thrust, the fuel injection of throttle valve. When the engine torsion is higher, it means that driver is speeding up. The acceleration can increase fuel consumption (0.024 l/km). Thus, avoiding the aggressive driving behavior can improve the fuel consumption.
2. Rotation per minute (RPM): the fuel consumption is higher when rotation is higher. The fuel consumption increases 0.002 l when rotation heightens. Thus, avoiding the frequent rapid acceleration can efficiently save the fuel consumption.
3. Speed: increasing speed can bring out shift frequency, acceleration processing and throttle valve. Following the result of analysis, the fuel consumption can increase 0.028 l when the speed goes up. Thus, it is useful to avoid frequent acceleration or high speed in order to reduce the fuel consumption.
4. Engine temperature: the engine temperature is higher when vehicles are in high speed operation. Thus, high speed operation must output higher fuel consumption. When engine temperature goes up, the fuel consumption can increase 0.225 liter per kilometer. In other words, there exists a relationship between the driving behavior, vehicle status and vehicle operation.
5. Trip distance: longer trip distance can bring out higher fuel consumption. The trip distance could directly affect the fuel consumption. As the result of regression analysis, the average fuel consumption per kilometer is 0.038 l.

5 Conclusion

This study provides academics and practitioners a macro view of the approaches for implementing green public transportation. Specifically, the proposed big data analytics can be used to evaluate bus-driving behavior. Future research could collect more bus OBD-II data to develop energy efficiency policies for public transportation. The continuing relevance of the proposed business analytics is expected in further studies, as there are many possible applications of research findings.

Our research aims at collecting the data from OBD II to display the relationship between driving behavior and fuel consumption. Through the data processing and data analysis, it can induce three attribution and 6 variables of fuel consumption. Using multi-regression analysis to find the statistical significance for driving behavior and fuel consumption is core value in the research. Thus, the result of analysis in the research can also discover that 5 variables, engine torsion, rotation, speed, engine temperature and trip distance, have a positive correlation and statistical significance with fuel consumption.

This research focus on the induction of fundamental elements of fuel consumption and linear regression method in order to find the correlation between driving behavior and fuel consumption, in which engine torsion, rotation and speed exist high correlation with drivers' behavior. In other words, reinforcing fleet management of eco-driving system can link to these relative variables. Our research first aims at analyzing the eco-driving of bus operator, it could be expanded to other transportation industries. For detail, it can compare the difference between the operators, the routes, the vehicles and the characteristics of drivers.

References

Alessandrini, A., Filippi, F., Orecchini, F., Ortenzi, F.: A new method for collecting vehicle behaviour in daily use for energy and environmental analysis. Proc. Inst. Mech. Eng. Part D **220**(11), 1527–1537 (2006)

Berry, I.M.: The effects of driving style and vehicle performance on the real-world fuel consumption of U.S. light-duty vehicles. Master's thesis, Mechanical Engineering, Virginia Polytechnic Institute and State University (2007)

Beusen, B., Broekx, S., Denys, T., Beckx, C., Degraeuwe, B., Gijsbers, M., Scheepers, K., Govaerts, L., Torfs, R., Panis, L.I.: Using on-board logging devices to study the longer-term impact of an eco-driving course. Transp. Res. Part D: Transp. Environ. **14**(7), 514–520 (2009)

Browne, G.: Design of experiments to determine factors contributing to fuel economy. Working paper (2008)

Ericsson, E.: Variability in urban driving patterns. Transp. Res. Part D **5**(5), 337–354 (2000)

Ericsson, E.: Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. Transp. Res. Part D **6**(5), 325–345 (2001)

Evans, L.: Driver behavior effects on fuel consumption in urban driving. Hum. Factors **21**(4), 389–398 (1979)

Hooker, J.N.: Optimal driving for single-vehicle fuel economy. Transp. Res. Part A **22**(3), 183–201 (1988)

Reiter, U.: Modeling the driving behavior influenced by information technologies. In: Highway Capacity and Level of Service, pp. 309–320 (1991)

Stæhr, J.: Økonomikørsel - kør grønt - for klimaets skyld (2013). http://www.oekonomikoersel.dk/. Accessed 9 February 2012

Waters, M.H.L., Laker, I.B.: Research on fuel conservation for cars. Report no. 921. Transportation Research Laboratory, Crowthorne (1980)

Evaluation of Traffic Network Performance Under Autonomous Vehicles with Intelligent Signal Control Policies

Li-Wen Chen(✉)

School of Economics and Management, MinJiang University, Fuzhou, Fujian 350108, China
lwchen2017@yahoo.com

Abstract. Intelligent Transportation Systems (ITS) focus on increasing the efficiency of existing surface transportation systems via the use of advanced computers, electronics, and communication technologies. A new era is emerging in terms of vehicle technology. With intelligent vehicles, no driver is needed to control the vehicle in the future. The advancement of vehicle technology has been able to provide autonomous vehicles in road traffic environment. Possible impacts of autonomous vehicles on density, speed, and flow, and vehicular movements on links and intersections might be affected as well. This research aims at developing an integrated simulation framework to investigate possible impacts of the autonomous/connected vehicles and to identify key issues on the impacts of autonomous/connected vehicles on road traffic, in terms of traffic flow and signal control policies. Speed-density relationships are assumed based on the autonomous vehicle's characteristics and three signal control policies are assumed to simulate network performance under different demand levels. Numerical experiments with different assumptions are conducted to evaluate system performance under autonomous vehicle environment.

Keywords: Automated vehicles · Traffic flows impact
Intelligent signal control

1 Introduction

Basic elements in road traffic include driver, vehicle and roadway. In the past, researchers on how to enhance communications among these three elements have been proposed and some of the idea have been implemented in real world, such as road guidance and electronic payment, etc. All these development are the success of Intelligent Transportation Systems (ITS), and ITS focus on increasing the efficiency of existing surface transportation systems through the use of advanced computers, electronics, and communication technologies [1–3].

A new era is emerging in terms of vehicle technology. With the intelligent of vehicles, no driver is needed to control the vehicle in the future. The advancement of vehicle technology has been able to provide autonomous vehicles in road traffic environment. With the advancement of vehicle technologies, applications of autonomous vehicles can be expected in a near future and field experiments are conducted

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 352–359, 2019.
https://doi.org/10.1007/978-3-030-04585-2_43

around the world. Recent studies report that the technology of vehicle-automation has become matured enough to enable the commercialization of autonomous vehicles, possibly by the 2020s [4]. Recent demonstrations of autonomous vehicles on the public roads also allude drastic changes in the auto industry [5]. In addition to autonomous vehicles, connected vehicle technologies provide the opportunity to create an interconnected network of moving vehicular units and stationary infrastructure units, in which individual vehicles can communicate with other vehicles (i.e., V2V communication) and other agents (e.g., a centralized traffic management center through V2I communication) [3].

This research proposes a research framework to investigate possible impacts of autonomous/connected vehicles by using mesoscopic simulation models with revised speed-density models to evaluate possible impact of autonomous vehicles. The proposed framework establishes the link between autonomous/connected vehicles with existing methodologies in traffic flow and management. The core is to introduce the concept of autonomous vehicle in traffic flow models and analyze possible impact on different levels of autonomous levels.

Two major considerations in the simulation are (1) how vehicles move on roadways and (2) how vehicles pass the intersections. Based on calibrated speed-density models, different speed-density relationships are assumed. Intersection control is classified into pre-timed signal control, actuated signal control, and intelligent intersection control. Numerical experiments are conducted in a subnetwork of Kaohsiung, Taiwan to reflect future application in urban cities. The proposed framework and associated modeling issues are described in Sect. 3. Numerical experiments based on a test network are discussed in Sect. 4, followed by a brief summary.

2 Literature Review

In 2016, NTHSA based on SAE (Society of Automotive Engineers) updates the levels of autonomous vehicles. There are 6 levels of automation, 0–2 levels involve human driver to control the vehicle, and levels 3–5 involve automated driving environment.

The operation of automated vehicles has the potential to substantially affect traffic safety, mobility, congestion, energy, and land use [1, 5, 6]. Most existing studies report such potential impacts of automated vehicles in rather speculative ways due to unknown human factors and unresolved legal issues attendant to the general-purpose operations of automated vehicles. Also, the extent to which automated vehicles benefit the society as a whole will also largely depend on their share in the total vehicle fleets [7, 8]. With new communication technologies, Shladover et al. [9] and Pinjari et al. [7] discussed that road capacity improvement shall be maximized if the technology of automated vehicles is combined with V2V (vehicle-to-vehicle) or V2I (vehicle-to-infrastructure) communication. Furthermore, ITF [10] pointed out that implementation of automated vehicles might even reduce required parking spaces themselves.

3 Research Methodology

Through extensive sensing and massive intelligence, autonomous vehicles possess the ability to drive without external assistance. Due to the advancement of vehicle technology, impacts to traffic networks need to be studied for traffic management. Possible impacts of autonomous/connected vehicle are summarized as follows: the average headway, speed-density relationship, saturation flow rates, possible increase in free flow speed, and intersection Control.

3.1 Impact of Autonomous Vehicles on Mixed Traffic Flows

The car-following models and traffic stream models have been developed and applied in a variety of applications to study traffic flows. Under autonomous vehicle, vehicle characteristics have been greatly enhanced and human role in driving is reducing. Therefore, new set of car-following and traffic stream models need to be re-examined and re-developed. This research focuses on how to model autonomous vehicles on a basis of traffic science theory with assumptions on speed-density relationships.

As illustrated in Fig. 1, the study starts with basic traffic flow theory, i.e. flow is equal to the produce of speed and density. New trends of vehicle technologies might change possible speed and density relationship. The boundary conditions are assumed based on different levels of vehicle automation and market penetration. Then, revised speed-density relationships are applied in the mesoscopic simulation, DynaTAIWAN [11] to evaluate possible impacts of autonomous/connected vehicles. Traffic flow distributions can thus be applied in designing traffic information and management

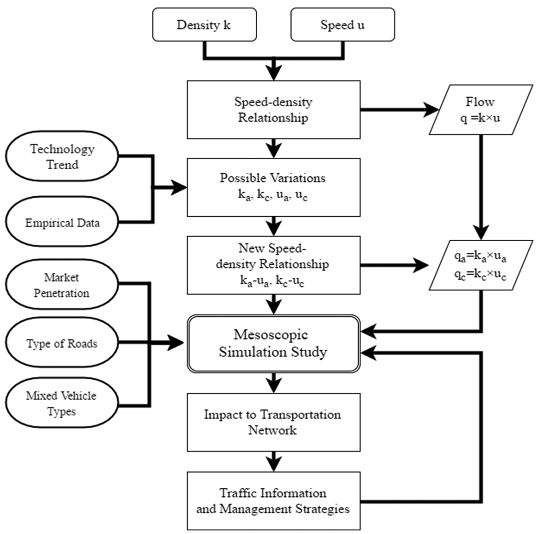


Fig. 1. A mesoscopic simulation framework for modeling autonomous vehicles

strategies. Scenarios are designed to reflect possible factors, such as market penetration of autonomous vehicles, types of roads, and mixed traffic flow conditions [12, 13]. Finally, network performance is simulated and evaluated.

Although there is still no data for traffic flow under autonomous and/or connected vehicles, some possible theoretical diagrams can still be assumed based on assumptions of vehicle technologies. For example, speed-density models for freeways and speed-density models for urban streets with the considerations of mixed traffic flow characteristics. A major determination of the impact of autonomous vehicles on traffic flow and urban mobility will be the extent to which these vehicles are adopted and accepted by users and the fraction of the total vehicle mix that they constitute. In Taiwan, applications of autonomous vehicles should first be deployed on freeway systems because of the relatively simple traffic stream characteristics.

3.2 Simulation

Vehicular movements in the traffic network are simulated in DynaTAIWAN through link movement and node transfer processes. The process of link movement moves vehicles on the link based on defined speed-density flow models and the process of node transfer process transfer vehicles from links to links based on predefined paths. In the link movement process, a modified Greenshields model, is assumed, as illustrated in the following equation, and adopted in DynaTAIWAN to simulate traffic movements. This revision adds a minimum speed in the model to reflect vehicular flow characteristics in a network. If no minimum speed is imposed, flow stops when density approaches jam density.

Within the DynaTAIWAN model, link density is calculated based on total number of vehicles. K_i^t, in terms of passenger car units (PCU), is calculated based on numbers of passenger car, truck, and motorcycle on link i and then the link speed V_i^t is obtained. Then, link speed is calculated based on the current density and given speed-density models. Based on the collected traffic flow data, speed-density models are calibrated. There are six types of models, including freeway/expressway, on-ramp/off-ramp, surface streets with different road types. Two models are described hereafter.

Under autonomous vehicle environment, driver response time can be significant reduced and acceleration/deceleration capabilities are improved. Based on the two characteristics, we can postulate that speed might be higher with the same density when compared with traditional traffic flows. Thus, speed-density models might have different shapes. Without real data, it is difficult to describe traffic flow models with autonomous vehicles. However, several assumptions can still be made to predict possible traffic flow behavior. In this research, we still assume traditional traffic flow behavior, such as Greenshield model, still hold, especially for boundary conditions. Higher speed can be achieved under same density situations. Based on these assumptions, we can assume possible traffic flow models by varying existing coefficients.

If we assume possible impact of autonomous vehicles come from higher speed and higher acceleration/deceleration capabilities, several possible models are generated by varying α coefficients. Models are illustrated in Fig. 2 and described as follows.

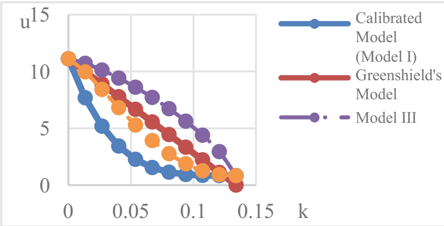


Fig. 2. Different speed-density models based on autonomous vehicle assumption

Freeway/expressway	Surface streets
Model I	
$V_i^t = (110 - V_i^0)\left(1 - \left(\frac{K_i^t}{164}\right)^{1.3}\right)^{5.12} + V_i^0 \quad (1)$	$V_i^t = (42.43 - V_i^0)(1 - \frac{K_i^t}{134})^{2.694} + V_i^0 \quad (2)$
Model II	
$V_i^t = (110 - V_i^0)\left(1 - \left(\frac{K_i^t}{164}\right)^{1.3}\right)^{3.072} + V_i^0 \quad (3)$	$V_i^t = (42.43 - V_i^0)(1 - \frac{K_i^t}{134})^{1.616} + V_i^0 \quad (4)$
Model III	
$V_i^t = (110 - V_i^0)\left(1 - \left(\frac{K_i^t}{164}\right)^{1.3}\right)^{1.024} + V_i^0 \quad (5)$	$V_i^t = (42.43 - V_i^0)(1 - \frac{K_i^t}{134})^{0.539} + V_i^0 \quad (6)$

In addition to the original speed-density models, two sets of models are experimented in this study. The coefficients are adjusted to reflect possible variations of speed-density models.

4 Numerical Experiments

The purposes of the experiment are (1) to observe how system performs under different assumptions on speed-density flow models and (2) to illustrate possible benefits from autonomous vehicles.

4.1 Network Descriptions

Numerical experiments are tested on district in Kaohsiung City shown in Fig. 3. The network consists of 27 traffic zones, 132 nodes and 363 links. The arcs consist of freeways, expressways and urban streets with real road characteristics. The traffic demands depending on the historical OD data provided the DynaTAIWAN to simulate the real network in the peak and off-peak traffic. There are also railway system and mass rapid transit in this area.

4.2 Experiment Setups

Three demand levels, low (94,980 vehicles), medium (140,900 vehicles), and high (203,500 vehicles), are designed to observe how in numerical analysis and possible variation of system performance in traffic networks with autonomous vehicles. In terms of signal control, three signal control policies, including pre-timed signal control, actuated signal control, and intelligent signal control, are designed to observe possible variations of traffic performance. With the applications of autonomous vehicles, the concept of signal control needs to be re-modelled in order to achieve fully benefits of autonomous vehicles. One of the possible approaches is "reservation", in which a vehicle approaching the intersection can request a block of space-time within which to traverse the intersection. Three major system performance observed include average travel time (in minutes), average travel distance (in kilometers), and average stopped time (in minutes).

4.3 Result Analysis

Results for three models under three demand levels are observed from the experiment. The variations of ATT and AST are depicted in Fig. 4. Some observations are discussed as follows:

(1) System performances in terms of ATT and AST increase exponentially with respect to number of vehicles, which confirms with a priori knowledge.
(2) The trends of ATT and AST exhibit with similar patterns. The results show that travel time is highly related with stopped time in the traffic network.
(3) System performances under different speed-density models are different. The different between Model I and III are significant, but not as in the differences between Model I and II. Models I and II provide similar patterns for speed estimation.
(4) With different assumptions of autonomous vehicles, speed-density models can be modelled as Models I, II, and III. However, the system performances generated through these models do not provide significant impact of traffic system. In order to evaluate the impact of autonomous vehicles, signal control needs to be specific considered and experimented.

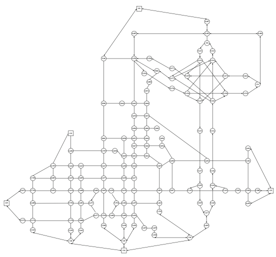


Fig. 3. The subnetwork of Kaohsiung City

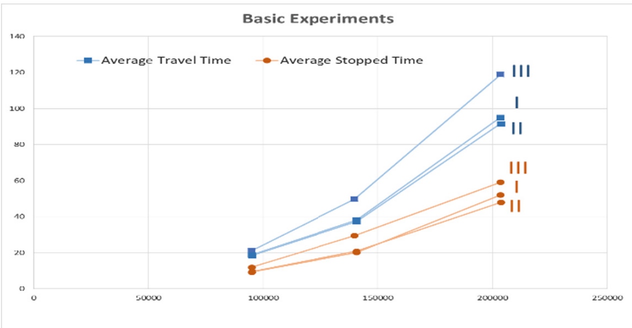


Fig. 4. Numerical results for three models under three demand levels

Two signal control strategies are experiments in numerical analysis. The first signal strategy is actuated signal control policy for two-way intersections and the second signal strategies assume intelligent control policy with "reservation" for passing intersections. For actuated signal control policy, minimum green time and maximum green time are assumed for all two-way intersections, and pre-timed signal control is applied for multi-phase controlled intersection. The second policy assumes that all vehicles can be coordinated and passed the intersection without further delay as long as there is enough capacity. The variations of ATT and AST are illustrated in Figs. 5 and 6.

Some observations are summarized as follows:

(1) Since high demand traffic conditions are simulated, actuated signal control perform only as good as pre-timed signal control. No significant benefit is observed; however, significant reductions of ATT and AST are observed under intelligent signal control policy.
(2) In terms of ATT, ATTs for pretimed signal policy and intelligent signal control policy are 89.93 min and 42.48 min, respectively. The results show that intersection delay might be the critical factor in traffic network. We can expect autonomous vehicle with connected capability might significant reduce intersection delay in traffic networks.
(3) The speed-density models I, II, and III provide similar trend in the experiments. The results show that speed-density relationships do not change vehicular flows patterns as long as the boundary conditions hold, i.e. jam density and free flow speed.
(4) The variations of ATT and AST also exhibit similar patterns.

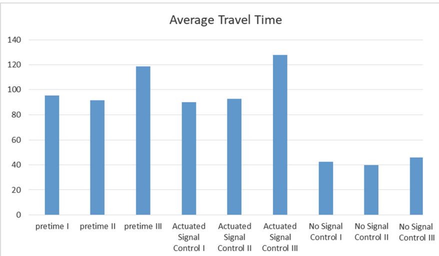


Fig. 5. Variations of ATT under three control policies

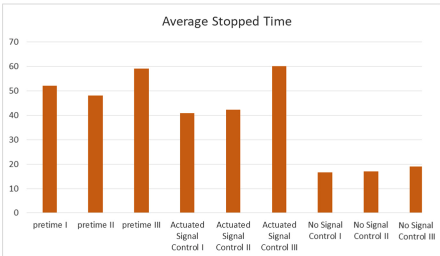


Fig. 6. Variations of AST under three control policies

5 Concluding Remarks

An integrated simulation framework is proposed to investigate possible impacts of the autonomous vehicles, in terms of traffic flow and system performance. Impacts of autonomous vehicles on density, speed, and flow are discussed through assumptions on vehicular movements on links. Speed-density relationships are assumed based on the autonomous vehicle's characteristics and mesoscopic simulation-assignment model is adopted to simulate network performance under different assumptions. Numerical experiments with different assumptions are conducted to evaluate system performance under autonomous vehicle environment.

Although speed-density relationships might vary with autonomous vehicles, little impacts of system performance in terms of reducing travel time are observed. On the contrary, intersection control play an important role to reduce travel time and stopped time. More research effort should direct to this direction in order to reduce congestion in urban cities. With more field experiments, traffic flow data can be collected and investigated in the future. Possible traffic flow theories can be further examined under mixed traffic flow characteristics.

References

1. Anderson, J.M., Kalra, N., Stanley, K.D., Sorensen, P., Samaras, C., Oluwatola, O.A.: Autonomous vehicle technology a guide for policymakers, p. 185. RAND Corporation, Santa Monica (2016)
2. Mahmassani, H.S.: Technological innovation and the future of urban personal travel. In: MOBILITY 2050: A Vision for Transportation Infrastructure, pp. 41–62 (2016)
3. Mahmassani, H.S.: 50th anniversary invited article—autonomous vehicles and connected vehicle systems: flow and operations considerations. Transp. Sci. **50**(4), 1140–1162 (2016)
4. Litman, T.: Autonomous Vehicle Implementation Predictions Implications for Transport Planning. Victoria Transport Policy Institute (2016)
5. Dokic, J., Müller, B., Meyer, G.: European roadmap smart systems for automated driving. In: European Technology Platform on Smart Systems Integration, p. 39 (2015)
6. Pinto, C.: How autonomous vehicle policy in California and Nevada addresses technological and non-technological liabilities. Intersect **5**(1), 1–16 (2012)
7. Pinjari, A.R., Augustin, B., Menon, N.: Highway capacity impacts of autonomous vehicles: an assessment. University of South Florida (2013)
8. Litman, T.: The new transportation planning paradigm. ITE J. **83**(6), 20–28 (2013). (www.ite.org)
9. Shladover, S., Su, D., Lu, X.: Impacts of cooperative adaptive cruise control on freeway traffic flow. Transp. Res. Record: J. Transp. Res. Board **2324**, 63–70 (2012)
10. ITF, Urban Mobility: System Upgrade, International Transport Forum and Corporate Partnership Board (2014). (www.internationaltransportforum.org)
11. Hu, T.Y., Tong, C.C., Liao, T.Y., Chen, L.W.: Dynamic route choice behaviour and simulation-based dynamic traffic assignment model for mixed traffic flows. KSCE J. Civ. Eng. **22**(2), 813–822 (2018). Online May 2017
12. Gartner N., Messer C., Rathi, A.: Traffic flow theory: a state of the art report, sponsored by TRB Committee A3A11 (2001). http://www.tfhrc.gov/its/tft/tft.htm
13. May, A.D.: Traffic Flow Fundamentals. Prentice-Hall, Upper Saddle River (1990)

Application of TRIZ to Determine Strategies for Transitioning from Third-Party to Fourth-Party Logistics in China

Ching-Kuei Kao[1(✉)], Mingwei Qian[2], Tang-Hsien Chang[3], and Peng-Jung Lin[4]

[1] Department of Logistics Management, Fujian University of Technology, No. 3, Xueyuan Road, University Town, Minhou, Fuzhou 350118, Fujian, China
ckkao@fjut.edu.cn

[2] Department of Logistics Management, Beijing Institute of Technology, No. 6, Jinfeng Road, Tangjiawan, Zhuhai 519088, Guangdong, China
524475151@qq.com

[3] Department of Transportation, Fujian University of Technology, No. 3, Xueyuan Road, University Town, Minhou, Fuzhou 350118, Fujian, China
thchang@fjut.edu.tw

[4] Department of Marketing, Beijing Institute of Technology, No. 6, Jinfeng Road, Tangjiawan, Zhuhai 519088, Guangdong, China
linpengjung@qq.com

Abstract. With the rapid development of global e-commerce, third-party logistics may be unable to adapt to the logistics environment under the development of e-commerce. The supply chain model of fourth-party logistics has gradually become a trend. In China, third-party logistics remains relatively mature. Various logistics enterprises are still comparatively unfamiliar with fourth-party logistics. This study compares the different points in executing third-party and fourth-party logistics to analyze the main problems faced by logistics enterprises when transitioning from third-party to fourth-party logistics. In the process of organizational change, if the organization has the capability to change management, it can appropriately improve the organization's cooperation and reduce the resistance of members. Therefore, this paper chooses "change management" as an important key capability to study the TRIZ strategy when China's third-party logistics is transformed into fourth-party logistics. On the basis of TRIZ strategies, this study provides reference for third-party enterprises that need transformation and help them to smoothly transform into fourth-party logistics.

Keywords: Fourth-party logistics · Third-party logistics · E-commerce · TRIZ

1 Introduction

Economic globalization and electronic information technology are gradually developing in depth, which renders the status and role of e-commerce increasingly prominent in the economic trade of many countries. E-commerce has become a trend in China's foreign trade development. The financial crisis that broke out in 2008 provided an opportunity

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 360–369, 2019.
https://doi.org/10.1007/978-3-030-04585-2_44

for developing cross-border e-commerce. Several enterprises took the opportunity to develop cross-border e-commerce, and e-commerce websites have been established to continuously explore the international market. Cross-border e-commerce enterprises with cross-border micro-transactions as their main business have been rapidly developed. At present, the marketing scale of global cross-border e-commerce market is 440 trillion US dollars, which accounts for 14% of the overall scale of e-commerce (Liu 2012). With the gradual development of cross-border e-commerce, this proportion is also increasing. As China's economic growth increases, the scale of cross-border e-commerce also becomes explosively upgraded. With the rapid development of e-commerce and cross-border e-commerce, China's third-party logistics (3PL) has gradually emerged and is unable meet the demand. Therefore, fourth-party logistics (4PL) is gradually arising in the field of logistics industry and becoming the focus. In China, the 4PL business model mainly includes Alibaba's rookie post and Feng-chao's automatic and intelligent delivery cabinet (Zhang and Shang 2015).

For cross-border e-commerce enterprises, the goods are mainly delivered to customers by air or maritime transportation. Third-party logistics enterprises can provide services for customers, such as custody, picking, inspection, packaging, reorganization, warehouse management and distribution. However, they may not have their own direct flights and must be outsourced to other shipping or airlines. Consequently, the logistics service efficiency and response speed are reduced. Compared to 3PL, 4PL is committed to providing complete supply chain logistics solutions (Ran 2004; Li 2006; Sun and Wang 2007). Therefore, 4PL has become a carrier of cross-border e-commerce companies for enabling them to smoothly expand foreign trade at a time when numerous enterprises are expanding their cross-border e-commerce business. At present, China's logistics is mainly based on 3PL. Although a prototype of 4PL is developed by certain 3PL, it is not full scale. In the development of e-commerce, a complex relationship exists between the users or providers of 3PL, so all partners in the supply chain are required to jointly maintain 3PL relationship. That is, 3PL enterprises that transition into 4PL is a general trend in the rapid development of e-commerce.

At present, the development of 3PL in China remains relatively mature. After the development of 3PL has matured (Hao 2015), each logistics enterprise wants to develop international logistics and seeks change. The quality of delivery and service quality have replaced the speed of logistics delivery and become the first important factor in user experience. The logistics problems of enterprises can be solved by features in 3PL, such as controlling logistics costs and providing efficient logistics. However, with the competition among enterprises and rapid development of e-commerce, customers require further services, such as e-procurement, order processing, and supply chain management. In cross-regional logistics, the limitations of 3PL are obvious. Consequently, developing into 4PL is imperative enterprises. Therefore, the transformation of China's 3PL enterprises into 4PL enterprises is an indispensable development if they want to secure a foothold in international logistics (Ouyang 2010).

China is a country with mature 3PL. However, 4PL remains a relatively unfamiliar model in various logistics enterprises, let alone transforming enterprises from 3PL to 4PL. Without a precedent for transformation, China's logistics enterprises may be unable to proceed smoothly or commence the process of transformation. Therefore, this study first integrates the definitions of 4PL by different studies. Then, it compares the different points in the execution of third- and 4PL to analyze the main problems faced by logistics enterprises when developing from third- to 4PL. The TRIZ method is used to build corresponding strategies for these main problems. According to these strategies, short, medium, and long-term strategies are compiled. This study provides strategic planning for each problem by solving the problems encountered in the process of transforming from 3PL to 4PL to provide further reference for third-party enterprises that need transformation and help them to smoothly transform into 4PL.

2 Literature Review

The definitions of 4PL vary across the perspective of different literature. By sorting out the definitions of 4PL, it is a supply chain model that integrates consulting, capital, and other resources based on 3PL to manage and apply these resources (Foster 1999; Han 2005; Liu 2005; Dollet and Diaz 2011). Moreover, Lieb and Kendrick (2003) proposed three modes of operation of 4PL, which are detailed in the following.

Synergy Plus Model
Although 4PL and 3PL jointly develop the market, 4PL can assist the efforts of 3PL to implement its plans and ideas, and provides services, such as supply chain strategy, technology, market entry capabilities, and project supervision, that 3PL fail to provide. In general, 4PL tends to work within 3PL enterprises, and ideas and strategies are implemented through 3PL as an effective implementer to achieve customer service goals. Fourth- and 3PL will adopt a strategic alliance or contractual mutual assistance approach to jointly develop the market (Lieb and Kendrick 2003; Wu and Liu 2008).

Solution Integrator Model
In this type, 4PL provides customers with a solution to operate and supervise a complete supply chain. 3PL and its resources and skills will comprehensively be managed by 4PL. As the core, 4PL can integrate the capabilities of various 3PL or service providers to fulfill customer needs to provide its customers with mediation planning to supervise and operate a complete supply chain (Lieb and Kendrick 2003; Wu and Liu 2008).

Industry Innovator Model
In the entire supply chain system, 4PL carries out the simultaneous integration and cooperation of the supply chain for different principals in multiple supply chains to obtain greater benefits for the members of the supply chain (Ran 2004). 4PL first integrates 3PL, then provides overall mediation planning to downstream customers. As the connection between upstream 3PL and downstream customer groups, the responsibility of 4PL is crucial. It can enhance the efficiency of the entire logistics industry through superior operational strategies, technology, and supply chain operations (Lieb and Kendrick 2003; Wu and Liu 2008).

Third-party logistics lacks the ability to operate across the supply chain and strategic expertise required to integrate supply chain processes. Fourth-party logistics can provide technological, warehousing, and transportation services and can effectively organize the best logistics providers in each segment to form an excellent supply chain management or logistics solution. The advantages of 4PL can perfectly break through the limitations of 3PL, and large-scale resource integration can truly achieve low cost and high efficiency.

3 Comparison Between Third-Party and Fourth-Party Logistics

An enterprise's supply chain gradually moves from insourcing to outsourcing. For example, logistics services in Europe have changed from market transactions to outsourcing to warehouses and contractors, then to the current 3PL and 4PL. From insourcing to outsourcing, the service process is provided by third-party goods, and 4PL will be the next major trend after 3PL. Therefore, a wide range of services should be provided for customers, and specialization in areas of these services must be established. These cues are an important process for the transformation from third- to 4PL. In addition to providing physical logistics services, increasing the number of customized services for customers, lengthening contact with customers, strengthening relationships, and improving the supply chain processes are necessary. To complete these processes, combining management consultancy, information technology, and 3PL and other skills is further necessary. Thus, 4PL is said to be a difficult supply chain strategy planner. The difference between fourth- and 3PL is that the former and its main customers frequently appear as joint ventures or strategic alliances. As an intermediary between

Table 1. Differences in economics and synergy between fourth-party and third-party logistics

	Third-party logistics	Fourth-party logistics
Service provider's participation in the overall supply chain	Entity activities and performance	Coordination and management of supply chain activities
Strength of knowledge provided by the service	Low (execution of standardization work)	High (flow of goods in the organization)
Contact window for service providers	Contact window for daily executive level staff and contract management	Prefers focused contact windows and higher level supply chain design and strategy collaboration
Performance of service provision	Measured by cost, quarterly productivity, and results	Provides a broader measure, such as customer service and measurement of supply chain strategies
Share strategic information with vendors	Enables awareness of changes in service levels, equipment, and other changes that have limited impact on logistics operations for 3PL	Broader and more comprehensive, including a list of customers and suppliers, service guidelines, and priorities
Reference	Guo and Wang (2005), Li (2006), Sun and Wang (2007)	Li (2006), Sun and Wang (2007)

customers and contact supply chain members, 4PL should assist customers and multiple logistics service providers in coordination, operation, and management.

At present, the performance of 3PL in China is considerable. However, performance in terms of customization and integration synergy remains insufficiently successful. Third-party logistics enterprises have gradually realized that they are facing new competitors who want to develop toward 4PL, such as transportation, cargo contracting, consulting, and software information enterprises and contract manufacturers (Wu and Liu 2008). Table 1 illustrates the economic and synergistic aspects of the transition from 3PL to 4PL. Based on the comparison in Table 1, 4PL pays more attention to the coordinated management of the process and customer participation mode than 3PL. That is, increased attention is paid to the operation process of the entire supply chain. In the supply chain, 4PL is more focused and flexible than 3PL on the coordinated management of the process, degree of service provision, and liaison with manufacturers.

Table 2 shows that the integration capability of 4PL is better than that of 3PL in terms of organization, positioning, services, and features. Compared with 3PL, 4PL involves multiple logistics service providers and involves all levels of work in the supply chain. Thus, the cross-functional integration capability is required for 4PL in multi-customer and supply chain management. In terms of logistics operations, 4PL not only possesses the characteristics of 3PL but also holds a high degree of management and responsibility that 3PL do not. Table 3 shows that the main difference between 3PL and 4PL is that the latter is relatively dynamic, which is reflected in its accountability for the operation of the entire supply chain. Third-party logistics is only a part of the supply chain and manages operations. However, 4PL holds responsibility for the operation and entire supply chain. Therefore, enterprises should have a strong ability to change management by being able change from partial management of the original supply chain to management of the entire supply chain as 3PL moves into 4PL.

Table 2. Comparison between third-party and fourth-party logistics

	Third-party logistics	Fourth-party logistics
Organization	Professional logistics service provider	An independent entity formed by a joint venture or long-term contract between the principal and partner
Positioning	Lack of cross-functional integration of operational capabilities in the supply chain	A supply chain integrator between the consignor and multiple logistics service providers
Service	Mainly for transportation and warehousing services	All supply chain level operations
Features	Integrate analytical management with technology to streamline unnecessary processes and reduce costs	In addition to integrating the resources and technologies of 3PL, it promotes a high degree of management and responsibility in the supply chain to create higher added value
Reference	Li (2007), Huang et al. (2008)	Cheng et al. (2008), Dollet and Diaz (2011)

Table 3. Main difference between providers of third-party and fourth-party logistics

	Third-party logistics provider	Fourth-party logistics provider
Characteristic	• Operational orientation • Implementer and performer • Typically more static information management Takes over one or part of the functioning of the supply chain • Managing other 3PL providers is unnecessary • Only one link in the supply chain	• Macroscopic • Integrator • Dynamic information management • Takes over the overall supply chain • Manages and coordinates all 3PL in the supply chain
Reference	Liu (2003)	Ran (2004)

Based on the table, 3PL tends to use contractual methods to provide specific logistics and supply chain activities to maximize the utility of their free asset use. However, its lack of ability can help customers manage the supply chain, and investment in R&D innovation is also low. Moreover, the development of 3PL is restricted because of limited skills in information technology and project management. As the 4PL provider of the new supply chain business model, it must have the leadership ability to gather the resources and skills of complementary organizations, such as 3PL, other consulting, and information technology enterprises to develop a competitive supply chain scheme for customers.

Through the common points and different points in the above comparisons, we can conclude that 3PL enterprises may soon become 4PL enterprises after improving their capabilities in the following areas, namely, capabilities in logistics operation, information technology application, and multi-customer and supply chain management, and change management (Table 4).

Table 4. Improvement capability and content during transformation

Improvement capability	Content
Logistics operation	Hold a higher level of logistics operation technology or integrated operation management capabilities based on the capabilities of 3PL
Information technology application	Integrated information to provide comprehensive information management
Multi-customer and supply chain management	Ability to manage and coordinate with the client or supply chain, and it can be flexibly managed throughout the supply chain
Change management	Capability for change management during transformation

4 Construction of TRIZ Strategies

Based on the results of the analysis in the previous section, the following key capabilities are obtained: “logistics operation,” “Information technology application,” “multi-customer and supply chain management,” and “change management.” In the process of organizational change, success or failure is related to the capability to change management. The biggest obstacle to any change comes from the slack of the organization and the resistance of the members. If the organization has the capability to change management, it can appropriately improve the organization’s cooperation and reduce the resistance of members. Therefore, this paper chooses “change management” as an important key capability to study the TRIZ strategy when China’s third-party logistics is transformed into fourth-party logistics.

For the analytical application of TRIZ, the following steps are proposed by Savransky (2008): (1) identifying problem, (2) determining improving and worsening parameters, (3) searching for the intentions of innovative principles, and (4) creating strategies based on the innovative principles according to contradictory matrix and 40 innovative principles in TRIZ theory. Therefore, the procedure of implementing TRIZ are follows.

Step 1. Identifying the capability

If an enterprise changes its business model, then a change in the entire operation of the enterprise will naturally follow. Therefore, although 3PL is excessive to 4PL, the enterprise must also change its management. This change can be a change in business strategy, a transfer of people, or a manner of retraining employees.

Step 2. Determining the improving and worsening parameters

Notably, a new operational mechanism should be introduced in the process of transforming management capabilities. Thus, the improving parameter is “10. Force”. When introducing or changing an operational mechanism, the employees should re-adapt in the workplace if mobility is given to the personnel. If the employees are retrained, then the employees are also familiarized with the new work. Therefore, the worsening parameter can be regarded as “36. Complexity of device.”

Step 3. Searching for intentions of innovative principles

According to improving parameter “10. Force” and worsening parameter “36. Complexity of device,” the innovation principles of the capability of change management that correspond to the contradiction matrix are “Principle 10. Preliminary Action,” “Principle 18. Mechanical Vibration,” “Principle 26. Copying,” and “Principle 35. Parameter Change.” From the analysis of various aspects, “Principle 10. Preliminary Action” and “Principle 18. Mechanical Vibration” are selected.

Step 4. Creating strategies based on innovative principles

The intention of “Principle 10. Preliminary Action” means pre-importing useful actions into an object or system. In change management, this intention can be reflected in the dialogue with employees before change management. Communicate fully with employees, and solicit willingness before mobilizing or retraining. If the employee does not accept or adapt to the transfer or training, then the passive work of the employee will naturally affect the operation of the enterprise.

The intention of "Principle 18. Mechanical Vibration" aims to use external components to cause oscillation or vibration. In this case, this intention can be understood as hiring a change management consultant or team for professional guidance. Thus, the enterprise must provide important information, such as the entire enterprise's operating structure, to the change management consultant or team and actively cooperate with their work. Another intention of "Principle 18. Mechanical Vibration" aims to increase frequency. This situation can be illustrated as an increase in communication. After major management changes, employees may be unfamiliar with their new job responsibilities. In this case, employees can ask questions, communicate with one another, and strive to maximize their work potential.

According to the above-mentioned statement, two new strategies of the important key capability "change management" are developed. Table 5 shows the corresponding improving and worsening parameters, the adopted innovative principles, and the developed strategies of the important key capability "change management."

Table 5. Strategies constructed by applying TRIZ approach

Key core capability	Improving parameter	Worsening parameter	Innovative principles	Strategy design
Capability of change management	10. Force	36. Complexity of device	10. Preliminary action	S8. Communication with various departments and collection of change opinions before the change in the operation
			18. Mechanical vibration	S9. Before the change, a professional team of consultants is hired to provide professional guidance; after the change, feedback is obtained through communication between employees

5 Conclusion

In response to the rapid development of China's e-commerce market demand, logistics enterprises should reflect on the changes in market demand, such that the establishment of goods circulation can further meet customer and merchant needs. At present, majority of the international cargo transportation and logistics centers are concentrated in Asia. If Asia or China is unable to meet the high-quality needs of customers and merchants, then the phenomenon will only give way to top international logistics enterprises in the US and Europe to stand firm in Asia or China. By then, the development of China's logistics industry will face major obstacles.

By comparing the difference between 3PL and 4PL, this study finds that 4PL is better than 3PL. The four key capabilities are obtained to provide logistics enterprises with reference for the requirements in transforming from 3PL to 4PL, namely, capabilities of logistics operation, information technology application, multi-customer and supply chain management, and change management. In the process of organizational change, if an organization has the ability to change management, it can appropriately improve organizational cooperation and reduce member resistance. Therefore, this paper chooses "change management" as an important key capability to study the TRIZ strategy when China's third-party logistics is transformed into fourth-party logistics.

On the basis of TRIZ strategies, the logistics enterprises should allocate the duty that each department should accomplish and strengthen the links and communication between various departments. The supply chain is a complete process, which requires close coordination of all links. If one department fails to cooperate, then the entire supply chain will be broken, such that it is unable to connect or even operate. In the process of change, introducing relevant logistics talents or training employees to understand and familiarize with the operation of 4PL may also be necessary.

References

Cheng, J.H., Chen, S.S., Chuang, Y.W.: A study of constructing fourth party logistics' selection criterion from supply chain integration and information technology perspectives – an application of fuzzy MCDM. Electron. Commer. Stud. **6**(4), 401–424 (2008)

Dollet, J.N., Diaz, A.: Supply chain orchestration for the luxury alcoholic beverage sector. IUP J. Supply Chain Manag. **13**(3), 42–65 (2011)

Foster, T.: 4PLs: the next generation for supply chain outsourcing? Logistics Manag. Distrib. Rep. **38**(4), 35 (1999)

Guo, F., Wang, H.M.: "4Rs" and marketing innovation of the third party logistics service provider. J. Northeast. Univ. (Soc. Sci.) **7**(5), 346–349 (2005)

Han, F.H.: Opportunities and challenges of fourth party logistics. Glob. Logistics Oper. J. **22**, 36–37 (2005)

Hao, H.X.: Analysis of China's third party logistics cost control. Prices Mon. **8**, 54–57 (2015)

Huang, F., Shi, X.L., Dai, X.Z.: Research on 3PL resource integration mode in Zhejiang Province. Logistics Technol. **27**(2), 25–28 (2008)

Li, B.L.: Research on the development model and strategy of the fourth party logistics in China. Sci. Technol. Inf. **1**, 133–134 (2006)

Li, S.M.: Research and Realization of the Collaborative Logistics Information Management System Based on the Third-Party Logistics, Sun Yat-sen University (2007)

Lieb, R.C., Kendrick, S.: The year 2002 survey: CEO perspectives on the current status and future prospects of the 3PL industry in the US. Transp. J. **43**(2), 5–16 (2003)

Liu, Q.Y.: Design of third party logistics management information system. J. China Railway **10**, 27–29 (2003)

Liu, Y.L.: The crisis of third-party logistics is also a turning point (2005)

Liu, J.: Application and Implementation of Forth Party Logistics in China's E-commerce, Harbin Institute of Technology (2012)

Ouyang, L.: Study on Transforming Model of Private Third-Party Logistics to Fourth-Party Logistics, Sun Yat-sen University (2010)

Ran, J.W.: The supply chain integration of the fourth party logistics. Commer. Res. **6**, 151–152 (2004)

Savransky, S.D.: Engineering of Creativity: Introduction to TRIZ Methodology of Inventive Problem Solving. Taylor & Francis, Milton Park (2008)

Sun, Y.B., Wang, D.P.: Research on operational modes of 4PL and its development trend in China. J. Beijing Technol. Bus. Univ. (Soc. Sci.) **22**(6), 85–90 (2007)
Wu, H., Liu, J.T.: Research on collaborative relationship between third party logistics and fourth party logistics. Sci. Technol. Manag. Res. **28**(1), 134–136 (2008)
Zhang, H.Y., Shang, X.: Analysis of last mile distribution modes for express industry: in the case of rookie station and fengchao delivery cabinet. Logistics Technol. **34**(11), 48–51 (2015)

Author Index

Y. Zhao et al. (Eds.): VTCA 2018, SIST 128, pp. 371–373, 2019.
https://doi.org/10.1007/978-3-030-04585-2