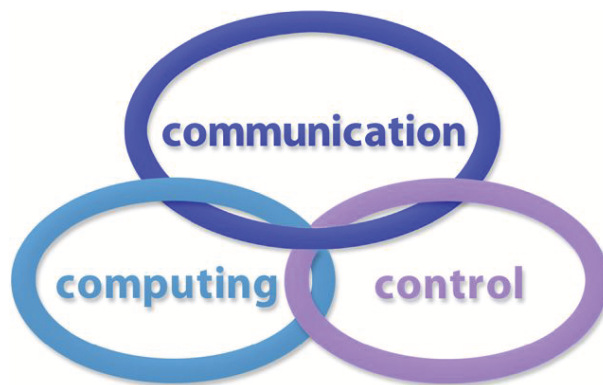


INTERNATIONAL JOURNAL
of
COMPUTERS, COMMUNICATIONS & CONTROL

ISSN 1841-9836

ISSN-L 1841-9836



A Bimonthly Journal
With Emphasis on the Integration of Three Technologies

Year: 2013 Volume: 8 Issue: 5 (October)

This journal is a member of, and subscribes to the principles of,
the Committee on Publication Ethics (COPE).



Agora University Editing House

CCC Publications

<http://univagora.ro/jour/index.php/ijccc/>

International Journal of Computers, Communications & Control



**EDITOR IN CHIEF:
Florin-Gheorghe Filip**

Member of the Romanian Academy
Romanian Academy, 125, Calea Victoriei
010071 Bucharest-1, Romania, ffilip@acad.ro

**ASSOCIATE EDITOR IN CHIEF:
Ioan Dzitac**

Aurel Vlaicu University of Arad, Romania
St. Elena Dragoi, 2, 310330 Arad, Romania
ioan.dzitac@uav.ro

&

Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea, Romania
rector@univagora.ro

**EXECUTIVE EDITOR:
Răzvan Andonie**

Central Washington University, USA
400 East University Way, Ellensburg, WA 98926, USA
andonie@cwu.edu

**MANAGING EDITOR DEPUTY MANAGING EDITOR
Mișu-Jan Manolescu Horea Oros**

Agora University of Oradea, Romania
Piata Tineretului, 8, 410526 Oradea
mmj@univagora.ro

University of Oradea, Romania
St. Universitatii 1, 410087, Oradea
horos@uoradea.ro

TECHNICAL SECRETARY

Cristian Dzitac
R & D Agora, Romania
rd.agora@univagora.ro

Emma Valeanu
R & D Agora, Romania
evaleanu@univagora.ro

EDITORIAL ADDRESS:

R&D Agora Ltd. / S.C. Cercetare Dezvoltare Agora S.R.L.
Piata Tineretului 8, Oradea, jud. Bihor, Romania, Zip Code 410526
Tel./ Fax: +40 359101032

E-mail: ijccc@univagora.ro, rd.agora@univagora.ro, ccc.journal@gmail.com
Journal website: <http://univagora.ro/jour/index.php/ijccc/>

International Journal of Computers, Communications & Control



EDITORIAL BOARD

Boldur E. Bărbat

Sibiu, Romania
bbarbat@gmail.com

Pierre Borne

Ecole Centrale de Lille
Cité Scientifique-BP 48
Villeneuve d'Ascq Cedex, F 59651, France
p.borne@ec-lille.fr

Ioan Buciu

University of Oradea
Universitatii, 1, Oradea, Romania
ibuciu@uoradea.ro

Hariton-Nicolae Costin

Faculty of Medical Bioengineering
Univ. of Medicine and Pharmacy, Iași
St. Universitatii No.16, 6600 Iași, Romania
hcostin@iit.tuiasi.ro

Petre Dini

Cisco
170 West Tasman Drive
San Jose, CA 95134, USA
pdini@cisco.com

Antonio Di Nola

Dept. of Mathematics and Information Sciences
Università degli Studi di Salerno
Salerno, Via Ponte Don Melillo 84084 Fisciano,
Italy
dinola@cds.unina.it

Ömer Egecioglu

Department of Computer Science
University of California
Santa Barbara, CA 93106-5110, U.S.A
omer@cs.ucsb.edu

Constantin Gaidric

Institute of Mathematics of
Moldavian Academy of Sciences
Kishinev, 277028, Academiei 5, Moldova
gaidric@math.md

Xiao-Shan Gao

Academy of Mathematics and System Sciences
Academia Sinica
Beijing 100080, China
xgao@mmrc.iss.ac.cn

Kaoru Hirota

Hirota Lab. Dept. C.I. & S.S.
Tokyo Institute of Technology
G3-49,4259 Nagatsuta, Midori-ku, 226-8502, Japan
hirota@hrt.dis.titech.ac.jp

George Metakides

University of Patras
University Campus
Patras 26 504, Greece
george@metakides.net

Ștefan I. Nitchi

Department of Economic Informatics
Babes Bolyai University, Cluj-Napoca, Romania
St. T. Mihali, Nr. 58-60, 400591, Cluj-Napoca
nitchi@econ.ubbcluj.ro

Shimon Y. Nof

School of Industrial Engineering
Purdue University
Grissom Hall, West Lafayette, IN 47907, U.S.A.
nof@purdue.edu

Stephan Olariu

Department of Computer Science
Old Dominion University
Norfolk, VA 23529-0162, U.S.A.
olariu@cs.odu.edu

Gheorghe Păun

Institute of Mathematics
of the Romanian Academy
Bucharest, PO Box 1-764, 70700, Romania
gpaun@us.es

Mario de J. Pérez Jiménez

Dept. of CS and Artificial Intelligence
University of Seville, Sevilla,
Avda. Reina Mercedes s/n, 41012, Spain
marper@us.es

Dana Petcu

Computer Science Department
Western University of Timisoara
V.Parvan 4, 300223 Timisoara, Romania
petcu@info.uvt.ro

Radu Popescu-Zeletin

Fraunhofer Institute for Open
Communication Systems
Technical University Berlin, Germany
rpz@cs.tu-berlin.de

Imre J. Rudas

Institute of Intelligent Engineering Systems
Budapest Tech
Budapest, Bécsi út 96/B, H-1034, Hungary
rudas@bmf.hu

Yong Shi

Research Center on Fictitious Economy
& Data Science
Chinese Academy of Sciences
Beijing 100190, China
yshi@gucas.ac.cn
and
College of Information Science & Technology
University of Nebraska at Omaha
Omaha, NE 68182, USA
yshi@unomaha.edu

Athanasios D. Styliadis

Alexander Institute of Technology
Agiou Panteleimona 24, 551 33
Thessaloniki, Greece
styl@it.teithe.gr

Gheorghe Tecuci

Learning Agents Center
George Mason University, USA
University Drive 4440, Fairfax VA 22030-4444
tecuci@gmu.edu

Horia-Nicolai Teodorescu

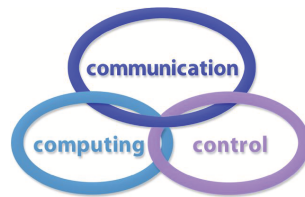
Faculty of Electronics and Telecommunications
Technical University "Gh. Asachi" Iasi
Iasi, Bd. Carol I 11, 700506, Romania
hteodor@etc.tuiasi.ro

Dan Tufiş

Research Institute for Artificial Intelligence
of the Romanian Academy
Bucharest, "13 Septembrie" 13, 050711, Romania
tufis@racai.ro

Lotfi A. Zadeh

Professor,
Graduate School,
Director,
Berkeley Initiative in Soft Computing (BISC)
Computer Science Division
Department of Electrical Engineering
& Computer Sciences
University of California Berkeley,
Berkeley, CA 94720-1776, USA
zadeh@eecs.berkeley.edu

**DATA FOR SUBSCRIBERS**

Supplier: Cercetare Dezvoltare Agora Srl (Research & Development Agora Ltd.)

Fiscal code: 24747462

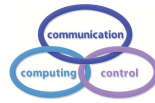
Headquarter: Oradea, Piata Tineretului Nr.8, Bihor, Romania, Zip code 410526

Bank: MILLENNIUM BANK, Bank address: Piata Unirii, str. Primariei, 2, Oradea, Romania

IBAN Account for EURO: RO73MILB000000000932235

SWIFT CODE (eq.BIC): MILBROBU

International Journal of Computers, Communications & Control



Short Description of IJCCC

Title of journal: International Journal of Computers, Communications & Control

Acronym: IJCCC

Abbreviated Journal Title: INT J COMPUT COMMUN

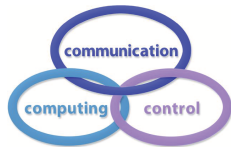
International Standard Serial Number: ISSN 1841-9836, ISSN-L 1841-9836

Publisher: CCC Publications - Agora University

Starting year of IJCCC: 2006

Founders of IJCCC: Ioan Dzitac, Florin Gheorghe Filip and Mişu-Jan Manolescu

Logo:



Publication frequency: Bimonthly: Issue 1 (February); Issue 2 (April); Issue 3 (June); Issue 4 (August); Issue 5 (October); Issue 6 (December).

Coverage:

- Beginning with Vol. 1 (2006), Supplementary issue: S, IJCCC is covered by Thomson Reuters - SCI Expanded and is indexed in ISI Web of Science.
- Journal Citation Reports(JCR)/Science Edition:
 - Impact factor (IF): JCR2009, IF=0.373; JCR2010, IF=0.650; JCR2011, IF=0.438; JCR2012, IF=0.441.
- Beginning with Vol. 2 (2007), No.1, IJCCC is covered in EBSCO.
- Beginning with Vol. 3 (2008), No.1, IJCCC, is covered in Scopus.

Scope: International Journal of Computers Communications & Control is directed to the international communities of scientific researchers in computer and control from the universities, research units and industry.

To differentiate from other similar journals, the editorial policy of IJCCC encourages the submission of scientific papers that focus on the integration of the 3 "C" (Computing, Communication, Control).

In particular the following topics are expected to be addressed by authors:

- Integrated solutions in computer-based control and communications;
- Computational intelligence methods (with particular emphasis on fuzzy logic-based methods, ANN, evolutionary computing, collective/swarm intelligence);
- Advanced decision support systems (with particular emphasis on the usage of combined solvers and/or web technologies).

Copyright © 2006-2013 by CCC Publications

Contents

Development an Adaptive Incremental Fuzzy PI Controller for a HVAC System J. Bai	654
Remarks on Interface Oriented Software Systems Modelling D. Bocu, R. Bocu	662
Estimation of the Text Skew in the Old Printed Documents D. Brodić, Č.A. Maluckov, L. Peng	673
SmartSteg: A New Android Based Steganography Application D. Bucerzan, C. Ratiu, M.J. Manolescu	681
Asymptotically Unbiased Estimator of the Informational Energy with kNN A. Cațaron, R. Andonie, Y. Chueh	689
Feature Clustering based MIM for a New Feature Extraction Method S. El Ferchichi, S. Zidi, K. Laabidi, M. Ksouri, S. Maouche	699
A Detailed Analysis of the GOOSE Message Structure in an IEC 61850 Standard-Based Substation Automation System C. Kriger, S. Behardien, J. Retonda-Modiya	708
Fast and Accurate Home Photo Categorization for Handheld Devices using MPEG-7 Descriptors B. Oh, J. Yu, J. Yang, J. Nang, S. Park	722
Feedback Linearization with Fuzzy Compensation for Uncertain Nonlinear Systems M.C. Tanaka, J.M.M. Fernandes, W.M. Bessa	736
Performance Analysis of Epidemic Routing in DTN with Overlapping Communities and Selfish Nodes Y. Wu, S. Deng, H. Huang, Y. Deng	744
A Quick Location Method for High Dynamic GNSS Receiver Based on Time Assistance P. Wu, S. Jing, W. Liu, F. Wang	754
Scalable Architecture for CPS: A Case Study of Small Autonomous Helicopter J. Yao, J. An, F. Hu	760
Distributed Genetic Algorithm for Disaster Relief Planning K. Zidi, F. Mguis, P. Borne, K. Ghedira	769
Author index	784

Development an Adaptive Incremental Fuzzy PI Controller for a HVAC System

J. Bai

Jianbo Bai

College of Mechanical and Electrical Engineering,
Hohai University,
Changzhou 213022, China,
bai_jianbo@hotmail.com

Abstract: This paper presents an adaptive incremental fuzzy PI controller (AIFPI) for a heating, ventilating, and air conditioning (HVAC) system capable of maintaining comfortable conditions under varying thermal loads. The HVAC system has two subsystems and is used to control indoor temperature and humidity in a thermal zone. As the system has strong-coupling and non-linear characteristics, fixed PI controllers have poor control performance and more energy consumption. Aiming to solve the problem, fuzzy control and PI control are combined together organically. In the proposed control scheme, the error of the system output and its derivative are taken as two parameters necessary to adapt the proportional (P) and integral (I) gains of the PI controller based on fuzzy reasoning according to practical control experiences. To evaluate the effectiveness of the proposed control methods in the HVAC system, it is compared with a fixed well-tuned PI controller. The results demonstrate that the AIFPI controller has more superior performance than the latter.

Keywords: HVAC system, adaptive control, fuzzy logic control, PI control.

1 Introduction

Commercial and industrial HVAC applications use electric and mechanical control system to maintain the desired temperature humidity level, and static pressure within a given area or zone. Good HVAC control schemes help reduce energy used and maintain occupant comfort. In spite of many advance in control theory, simply controllers of PI/PID type are still widely used in the majority of HVAC control loops [1]. There are three common methods for determine "good" values for the gain, integral time constant and derivative time constant of a PI/PID controller: manual tuning, auto tuning and adaptive control method [2]. For thermal load disturbances, variation of fluid flowrate, heat exchangers fouling or wear on valves, most HVAC systems have nonlinear, strong-coupling and time-varying dynamics [3]. A problem using conventional PI controllers in HVAC control systems is that control performance varies as conditions change and loops may become sluggish or oscillatory at certain times [2].

According to Åström, a fixed gain controller should be used for systems with constant dynamics, and adaptive control methods should be used for processes with time varying dynamics [4]. Consequently, we should use adaptive control methods in the HVAC industry. Some work has already been done in this area. For example, the development of control strategies for improving the performance of PID controllers using self tuning and adaptive control PI control techniques for HVAC systems has been studied in recently years [5]. The controllers normally have two parts: online identification and controller tuning. And the Recursive Least Square (RLS) method is commonly used by the identification part. However, it has been reported that unmodeled process disturbances and actuator hysteresis limit the effectiveness of the RLS [6]. Then, the traditional adaptive control methods based on the identification have limitations in HVAC control applications, such as absence of robustness.

Intelligent adaptive control using AI (Artificial Intelligence) techniques don't need the identification procedure and can also adapt to the various situations in real time [7]. It has both adaptability and robustness characteristics. Therefore, the adaptive control based on AI techniques has acquired more attention in application to HAVC systems since the end of last century [8]. Zaheer-Uddin [9] has utilized a Neuro-PID algorithm for tracking control of a discharge air temperature (DAT). The ANN (Artificial Neural Network) is used to calculate the gain coefficient of PID controller. Results show that the controller is able to track set point trajectories efficiently in the presence of disturbances. Seem [10] has described a method for automatically adjusting the gain and integral time of PI controllers based upon patterns that characterize the closed-loop response for HVAC systems. This new pattern recognition adaptive controller (PRAC) is easy to use and can provides near-optimal performance for a range of systems and noise levels.

Fuzzy logic control, also a method of AI techniques, is proposed by Zadeh in 1973 and has been widely applied to industry areas [11]. It presents a good tool to deal with complicated, non-linear and time-variant systems. To utilize the advantages of fuzzy control and the existing PI controllers in HVAC control systems, we design an adaptive incremental fuzzy PI (AIFPI) controller, within which the parameters of the PI controllers can be updated online as a function of operational conditions to adapt to various load disturbances in HVAC systems.

2 HVAC System and System Models

A single thermal zone HVAC system for control analysis is considered [12]. The schematic diagram of the HAVC system and its control system is shown in Fig.1. The HVAC system consists of the following main components: a cooling coil, a supply air fan, a thermal zone, connecting ductwork, water pipe, and filter. The control system includes temperature sensors, humidity sensors, a Variable Frequency Drive (VFD), a water valve and a centralized controller, etc.

The basic operation mode of the HVAC system is considered as follows: initially, 25% fresh air enters the system and is mixed with 75% of the recirculated air, the remained air is exhausted. Then, the mixed air passed through the cooling coil where it is conditioned according to the thermal zone designed temperature. Next, the conditioned air is delivered by a supply fan and the flow rate of the air is adjusted to maintain the thermal zone humidity. Furthermore, the supply air enters thermal space to offset sensible and latent loads acting on the system. Finally, the air in the room space is recirculated and the remained air is exhausted to the outside environment. During the operation, the flow rate of the water and the air is modulated to maintain the setting points of the temperature and the humidity in the thermal space.

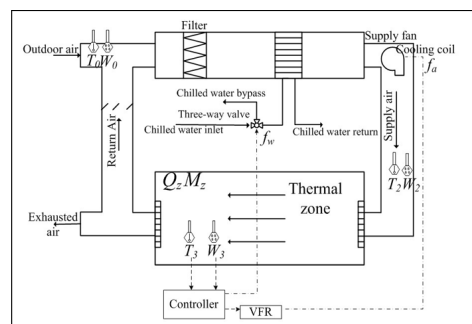


Figure 1: Schematic diagram of the HVAC system and its control system

According to the energy and mass balance of the HVAC system, the differential equations of the system mathematical model can be described as follows [12]:

$$\begin{cases} \frac{dW_3}{dt} = \frac{(W_2 - W_3)f_a}{V_z} + \frac{M_z}{\rho_a V_z} \\ \frac{dT_3}{dt} = \frac{(T_3 - T_2)f_a}{V_{he}} + \frac{0.25(T_0 - T_3)f_a}{V_{he}} - \frac{h_w(0.25W_0 + 0.75W_3 - W_2)f_a}{C_{pa}V_{he}} \\ \frac{dT_3}{dt} = \frac{(T_2 - T_3)f_a}{V_z} - \frac{h_{fg}(W_2 - W_3)f_a}{C_{pa}V_z} + \frac{1}{\rho_a V_z C_{pa}}(Q_z - h_{fg}M_z) \end{cases} \quad (1)$$

where C_{pa} is the specific heat of air (1.004kJ/kg.K); C_{pw} is the specific heat of water (4.183kJ/kg.K); f_a is the flow rate of air(m³/s); f_w is the flow rate of water(m³/s); h_w is the enthalpy of liquid of water(790.84kJ/kg); h_{fg} is the enthalpy of water vapor(2500.45kJ/kg); M_z is the moisture load (0.021kg/s); Q_z is the sensible heat load (84.93kJ/s); T_0 is the temperature of outside air (K); T_2 is the temperature of supply air (K); T_3 is the controlled temperature of thermal space(K); V_{he} is the volume of cooling coil (1.72m³); V_z is the volume of room space (1655.11m³); W_0 is the humidity of outside air (0.018kg/kg); W_2 is the humidity of supply air(0.007kg/kg); W_3 is the controlled humidity of thermal space (kg/kg); ρ_a is the density of air (1.185kg/m³); ρ_w is the density of Water(1000kg/m³).

3 Design of an Adaptive Increment Fuzzy PI Controller

Fig.2 shows the schematic diagram of the proposed controller and its application to the HVAC system. The control system is composed of two control loops. One is to control the indoor humidity (W_3) by regulating the frequency of the supply fan. The other is to control the indoor temperature (T_3) by regulating the opening of the bypass valve. Due to the complex characteristics of the HVAC system, it is difficult for fixed PI controllers to maintain the indoor environment and keep comfort situations

Fig.3 shows the infrastructure of the AIFPI controller. The control scheme is mainly composed of two parts: Fuzzy controller and PI controller, the increment of proportional and integral gains of the PI controller can be adjusted by the Fuzzy controller in real-time according to the error of the system output and the change of the error.

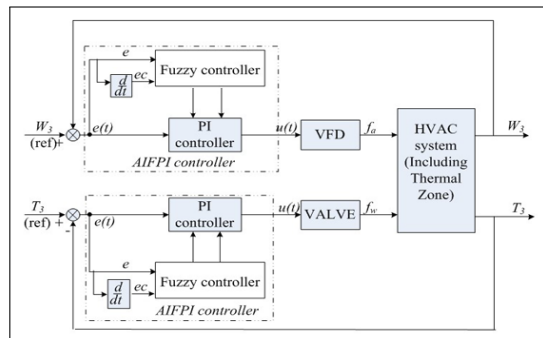


Figure 2: Schematic diagram of the AIFPI controller for the HVAC system

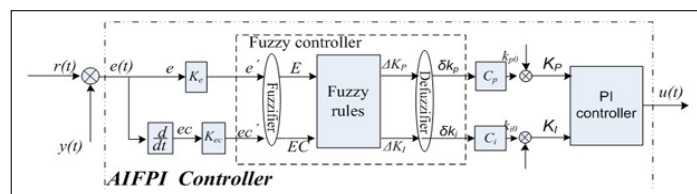


Figure 3: Infrastructure of the AIFPI controller

In Fig.3, $r(t)$ is the desired setpoint for process output and $y(t)$ is real process output; $u(t)$ is the fuzzy PI controller output; e is the error at sampling time t defined as $e = r(t) - y(t)$; ec is the change of the error; K_e and K_{ec} are input scaling factors of the fuzzy controller; Multiplied by K_e and K_{ec} separately, e' and ec' are the input variables of the fuzzy controller, which can be regarded as equivalent to e and ec . C_P and C_I are output scaling factors of the fuzzy controller; k_{p0} and k_{i0} represent initial proportional and integral gains of the PI controller; E and EC are Fuzzy sets of e' and ec' separately; ΔKP and ΔKI are fuzzy sets of δk_p and δk_i . They are the increment of proportional and integral gains of the PI controller. And K_P and K_I are proportional and integral gains of the PI controller. As the kernel of the fuzzy controller, the fuzzy rules are composed of the generalized forms "if-then" to describe the control policy and can be represented as follows.

$R^{(n)}$: If z Then $\{ \delta k_p \text{ is } \Delta KP_i^{(n)} \text{ and } \delta k_i \text{ is } \Delta KI_i^{(n)} \} \ i=1, \dots, m$;
 where $E_i^{(n)}, EC_j^{(n)}, \Delta KP_i^{(n)}$ and $\Delta KI_i^{(n)}$ are linguistic terms of $E, EC, \Delta KP$ and ΔKI .

According to the Mamdani inference rule, the fuzzy vector-matrix composition relation of ΔKP and ΔKI can be described as in follows:

$$\begin{cases} R_P = R_1 \cup R_2 \dots \cup R_n = \bigcup_{m=1}^n R_i \ (m = 1, \dots, 49) \\ R_I = R_1 \cup R_2 \dots \cup R_n = \bigcup_{m=1}^n R_i \ (m = 1, \dots, 49) \end{cases} \quad (2)$$

Then, the fuzzy rule bases for determination of ΔKP and ΔKI can be presented as

$$\begin{cases} \Delta KP = (E \times EC) \cdot R_p, \\ \Delta KI = (E \times EC) \cdot R_i, \end{cases} \quad (3)$$

As shown in Fig.3, the proposed fuzzy PI controller can be formulated as

$$\begin{cases} K_P = k_{p0} + \delta k_p \cdot C_p \\ K_I = k_{i0} + \delta k_i \cdot C_i \end{cases} \quad (4)$$

where the output crisp values, δk_p and δk_i can be calculated by the center of gravity methods. Thus, the output of the fuzzy PI controller can be given by

$$u(t) = K_P e(t) + K_I \int_0^t e(t) dt \quad (5)$$

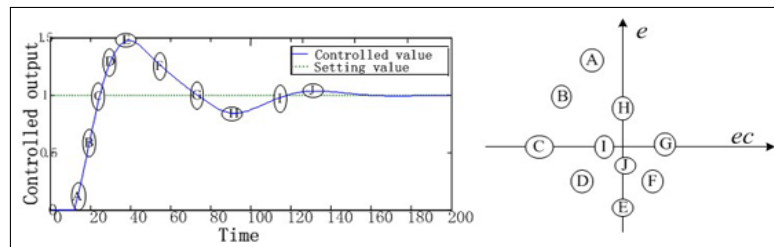


Figure 4: Infrastructure of the AIFPI controller

Fig.4 shows the diagram of controlled output regions at a step response and these regions' positions in a rectangular coordinate. The controlled output can be divided into ten regions. In the control scheme, the variation of incremental parameters of the PI controllers is relevant with the e and ec . And each of input and output variables are fuzzified using seven fuzzy sets. The

linguistic terms of the fuzzy sets are negative big (NB), negative medium (NM), negative small (NS), zero (ZO), positive small (PS), positive medium (PM), positive big (PB). The shape of the fuzzy sets, i.e. member functions, is chosen to be either triangle or Z-shape curve ones. The relation between ΔKP , ΔKI and E , EC can be summarized as the fuzzy rules in Table1, where the content in the first and second brackets represents the ΔKP and ΔKI separately.

Table 1 Fuzzy rule base for determination of ΔKP , ΔKI based on E and EC

E	EC						
	PB	PM	PS	ZE	NS	NM	NB
PB	(NB)(PB)	(NB)(PB)	(NM)(PM)	(NM)(PM)	(NM)(PS)	(Z)(Z)	(Z)(Z)
PM	(NB)(PB)	(NM)(PB)	(NM)(PM)	(NM)(PS)	(NS)(PS)	(Z)(Z)	(PS)(Z)
PS	(NM)(PB)	(NM)(PM)	(NS)(PS)	(NS)(PS)	(Z)(Z)	(PS)(NM)	(PS)(NS)
Z	(NM)(PM)	(NM)(PM)	(NS)(PS)	(Z)(Z)	(PS)(NS)	(PM)(NM)	(PM)(NM)
NS	(NS)(PS)	(NS)(PS)	(Z)(Z)	(PS)(NS)	(PM)(NS)	(PM)(NM)	(PM)(NB)
NM	(NS)(Z)	(Z)(Z)	(PS)(NS)	(PS)(NS)	(PM)(NM)	(PB)(NB)	(PB)(NB)
NB	(Z)(Z)	(Z)(Z)	(PS)(NS)	(PM)(NM)	(PM)(NM)	(PB)(NB)	(PB)(NB)

4 Application of the AIFPI Controller to the HVAC System

The AIFPI controller is applied to the HVAC system described in section 2. In the control system, there are two control loops to maintain the indoor temperature and humidity separately. There are hence two controllers in the control scheme (Fig.2). They have the same member functions of E , EC , ΔKP and ΔKI depicted in Fig.5. However, other parameters of the two AIFPI controllers, such as K_e , K_{ec} , C_p and C_i have different values to achieve good performance. From Fig.5, it can be found that the input variables of the fuzzy controller in the proposed control

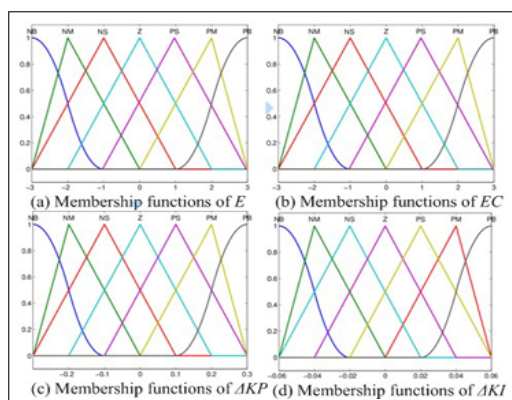


Figure 5: Member functions of the proposed AIFPI controller

scheme, e' and ec' , which are equivalent to the error e and the change of error ec are all quantized inside $[-3, 3]$. The output of the fuzzy controller, δk_p and δk_i , which are incremental proportional

and integral gains of the PI controller are quantized inside $[-0.06, 0.06]$ and $[-0.3, 0.3]$ separately. A proper choice of input and output scaling factors K_e , K_{ec} , C_p and C_i is important for the AIFPI controller to achieve good performance of the two control loops. After plenty of trials, the values of K_e , K_{ec} , C_p and C_i for controlling the indoor temperature are set as 6.5, 3, 1.6 and 10. And that for controlling the indoor humidity are 2.2, 4, 1.5 and 10. To evaluate the effectiveness of the proposed AIFPI controller, its control performance is compared with a fixed well-tuned PI controller, which adopts the PI/PID tuning rules proposed by Wang [13]. It has been proved that the auto-tuning controller using the PI/PID tuning rules has good performance for HVAC systems [14]. Their behavior is all evaluated with MATLAB/Simulink simulation using the HVAC system model described by Eq.(1) and predefined load disturbances, which are used to simulate non-linear, strong coupling and time-variable characteristics of the HVAC system.

(1) Comparison with different moisture load in the HVAC system

Fig.6 shows the controlled temperature and humidity subjected to different moisture load during the simulation. In the beginning, the moisture load is 0.021kg/s, the setting points of the humidity and the temperature are 0.009kg/kg and 298.15K. At the 400th sampling time, the moisture load is improved to 0.031kg/s to evaluate the control performance of the proposed controller. As shown in Fig.6, it can be found that the variation of the load disturbances is one of the important factors which will influence the stability of the controlled loops in the HVAC system. Good controllers can recover the stability and accuracy of the controlled variable quickly. It can be seen that the proposed controller achieves more superior performance than the well-tuned PI controller considering the transition time and the overshoot subjected to different moisture load disturbances. Fig.7 and Fig.8 also present the variation of the proportional and integral gains of the proposed controller during the simulation. It can be also found that the proportional and integral gains can adapt to the different load disturbances and attain new stable status when offsetting them.

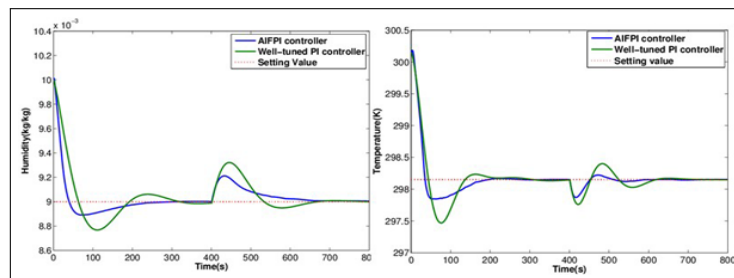


Figure 6: Controlled humidity and temperature with moisture load disturbances

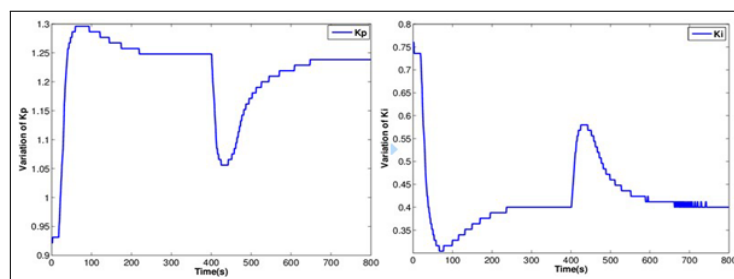


Figure 7: Variation of K_p and K_i of the AIFPI controller for controlling the humidity

(2) Comparison with different cooling load in the HVAC system

In HVAC systems, cooling load is changing all the time during a day. The variation of the cooling

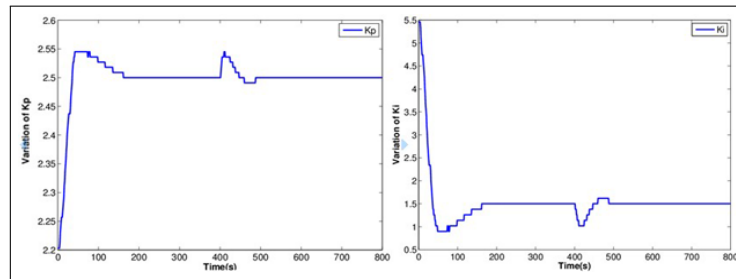


Figure 8: Variation of K_p and K_i of the AIFPI controller for controlling the temperature

load will lead to oscillation and drift of the controller variable, the indoor comfort then will be reduced. In order to evaluate the control performance subjected to cooling load disturbances for the proposed controller followed situations were simulated. In the beginning, the cooling load in the thermal zone is set as 84.93kJ/s, and the setting point of the humidity and the temperature is 0.009 kg/kg and 298.15K. At the 400th sampling step, the cooling load is increased by 40kJ/s.

Fig.9 shows the temperature and the humidity response of the two control loops. It can be seen that the proposed controller has less overshoot and quicker response speed than the well-tuned PI controller. It can be also found that the humidity will not be influenced subjected to the cooling load in the HVAC system. The variation of the proportional and integral gains of the controller controlling the temperature is shown in Fig.10. It can be concluded that the proposed controller has good adapt ability and control performance subjected to the cooling load disturbances.

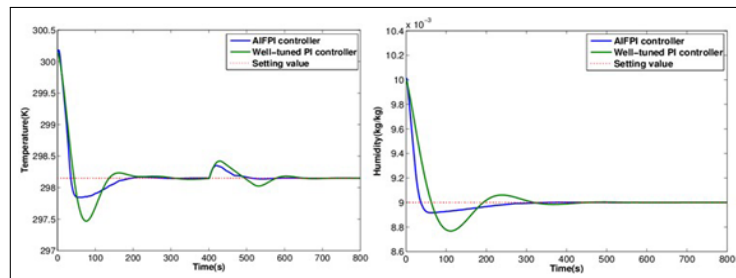


Figure 9: Controlled humidity and temperature with different cooling load disturbances

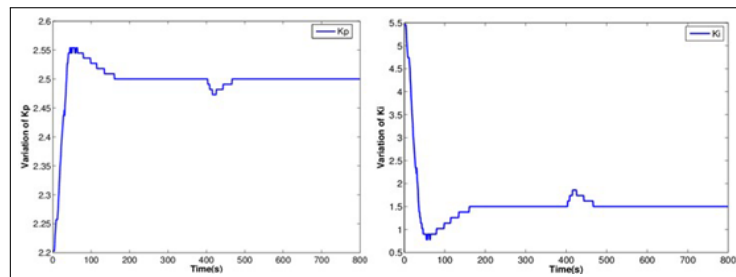


Figure 10: Variation of K_p and K_i of the AIFPI controller for controlling the indoor temperature

5 Conclusion

In this paper, an AIFPI controller is developed and applied to a HVAC system, which has two subsystems and is used to maintain temperature and humidity in a thermal zone. Taking full advantage of fuzzy logic control and PI control together, the proposed AIFPI controller uses Fuzzy logic to supervise PI controller parameters. In the control scheme, the incremental parameters of a PI controller are updated online as a fuzzy function of the operating conditions to improve the behavior of classical fixed PI controllers. The results demonstrate the AIFPI controller has good adaptability to the non-linear, strong coupling characteristics of the HVAC system. It performs significantly better control performance than the well-tuned PI controller considering response time and overshoot subjected to different moisture load disturbances or cooling load disturbances during the simulation. The AIFPI controller can be widely used in the HVAC industry.

Bibliography

- [1] Underwood, C.P., *HVAC control systems: modeling, analysis and design*, London and New York, E & FN Spon, 1999.
- [2] Salsbury, T., A survey of control technologies in the building automation industry, *Proc. of the 16th IFAC World Congress*, 331–341, 2005.
- [3] Levine, W.S., *Control System Applications*, Boca Raton, CRC Press, 121-123, 2000.
- [4] Astrom, K.J., Wittenmark, B., *Adaptive control*, Reading, Addison-Wesley, 1995.
- [5] Bai, J., Zhang, X., A new adaptive PI controller and its application in HVAC systems, *Energy Conversion and Management*, 48(2): 1043-1054, 2007.
- [6] Nesler, C.G., Automated Controller Tuning for HVAC Applications, *ASHRAE Trans.*, 92(2B): 189-201, 1986.
- [7] White, D.A., Sofge, D.A., *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, Van Nostrand, Reinhold Comp., 1992.
- [8] Ferreira, P.M., Ruano, A.E., Silva S., Neural networks based predictive control for thermal comfort and energy savings in public buildings, *Energy and Buildings*, 55(0): 238-251, 2012.
- [9] Zaheer-uddin, M., Tudoroiu, N., Neuro-PID tracking control of a discharge airtemperature system, *Energy Conversion and Management*, 45(15-16): 2405-2415, 2004.
- [10] Seem, J.E., Haugstad, H.J., Field and Laboratory Results for a New Pattern Recognition Adaptive Controller, *Proc. of Clima 2000 Conference*, 96-100, 2000.
- [11] Chang, S.S.L., Zadeh, L.A., On fuzzy mapping and control. Systems, Man and Cybernetics, *IEEE Trans. on*, (1): 30-34, 1972.
- [12] Arguello-Serrano, B., Velez-Reyes, M., Nonlinear control of a heating, ventilating, and air conditioning system with thermal load estimation, *IEEE Transactions on Control Systems Technology*, 7(1): 56-63, 1999.
- [13] Wang, Q.G., Lee T.H., Fung, H.W. et al., PID tuning for improved performance, *IEEE Trans. on Control Syst. Tech.*, 7(2): 457-465, 1999.
- [14] Bi, Q., Cai W.J., Wang, Q.G. et al., Advanced controller auto-tuning and its application in HVAC systems, *Control Eng. Practice*, 8(6): 633-644, 2000.

Remarks on Interface Oriented Software Systems Modelling

D. Bocu, R. Bocu

Dorin Bocu, Răzvan Bocu*

Department of Mathematics and Computer Science
Transilvania University of Brasov
Romania, 500036 Brasov
d.bocu@unitbv.ro

*Corresponding author: razvan.bocu@unitbv.ro

Abstract: This paper explores the foundations regarding the systematic usage of the concept of interface in order to sketch a methodological approach, in which the fundamental perspectives that guide the abstracting of a software system solution (referred to as UP, SP and BP in the paper) are unified, with the goal to optimally derive the behaviour of the system from its structure. Moreover, this is very useful for opening new avenues in order to address the shortcomings that are provoked by changes, considering a software system that is conceived at an industrial scale.

Keywords: user perspective, structural perspective, behavioural perspective, interface.

1 Introduction

Have all the issues that impede the realization of a quality software system been solved? We can't even speak about something like this. This is especially true nowadays, when the topological diversification of software systems has gone so far [2]. There are two reasons that will prevent this process from ceasing and will determine further amplification of it:

- An increasing number of the human activities are optimized or taken over by IT systems.
- The technologies that underline the creation of the IT systems bear an overwhelming strategic repositioning on the map that pictures the latest developments in the field.

In these conditions, the endeavour to inventorize the already existent paradigms and technologies becomes increasingly difficult. This process should have the essential goal of simplifying the choice of the paradigms and technologies that are necessary in order to roll out an IT project [3]. In this paper, we study several aspects that are related to the possibility of modeling the solution of a software system, while easing the natural demarche through which the switch from structure to behaviour is accomplished [7].

Although experts are largely unanimous regarding the qualitative determination of a system's behaviour by its structure [4], the experts themselves do not give up the search towards the identification of some new thought schemes concerning the relation between structure and behaviour in the process of a software system solution's modelling [5].

2 A Short Account on the State of The Art

Naturally, we can position ourselves far away from the beginnings regarding the technologies that are used in order to model the solution of a software system. Using a simplified and non-exhaustive approach, we can observe the following:

- In order to roll out the analysis activity in relation to an information system at a reasonable quality level, there are standard investigative methods and languages, with which the analysis results can be represented. The endeavour to optimize the streams of an information system involves the necessity to first realize a precise map of these streams. This is the main goal of the analysis activity. Naturally, considering a complementary layer, the analysis process is also intended to discover the requirements of the people that sustain the activity of the software system, regardless of the executive or managerial status of those who are involved [2]. We can offer examples of tools, with which it is possible to rationalize the activity of analysis: flow diagrams, Gantt diagrams, PERT diagrams, flow and control diagrams, UML and BPMN, etc. [1]. The technologies that feature footnotes represent two state-of-the-art approaches that aspire to the status of universal tools for the modelling activity [3].
- The progresses that have been achieved towards the rationalization of the modelling process are remarkable. It can be stated that in the war of methodologies an armistice has been reached, around the idea according to which the ideal solution for the specification of a modelling system is represented by the adoption of some rigour exposure formulae that is associated to a modelling process with the help of a notation that cautiously combines the formal correctness with the visual ingredients. It is a beneficial armistice both for developers and for the CASE tools developers.
- Considering the previous remarks, can it be stated that the modelling activity is free from syncope? It seems to be accurate to answer no. The modelling has continuously progressed, but it still faces open problems. As we have already mentioned, one of these problems is approached in this paper: the problem concerning the relation between the structure and the behaviour. More precisely, the idea that is highlighted is, briefly: how should the modelling activity be approached in order to ensure a reliable and flexible passage from the structure of a software system to its behaviour.

The structure of a software system, considering different levels of abstractization, represents a potential that has to be efficiently valued considering every phase of a software system solution's maturation [8].

If we do not work efficiently, we'll have to schedule additional iterations, which correct the modelling errors or eliminate the unintentional shortcomings of the modelling demarche [6]. What can we undertake in order to be more efficient in the modelling activity? An attempt to answer this question can be found in the following sections of this paper.

3 The Historic Premises of a Quality Modelling Endeavour

It has already been mentioned in Section 2 that specialists have constantly done their best in order to streamline the modelling of a software system [9]. With the risk to alienate part of the specialists that have already studied the same topic, it can be stated that the summary of these efforts can be expressed as follows:

From the structuring activity, using a methodic abstractization, towards object orientation. The structuring activity emphasized and sometimes turned into a fetish the necessity to structure data and the operations, following specific rules, with the goal to realize reliable and easy-to-maintain software systems.

During the era when the structuring activities used to prevail, mapping the solution domain on the problem domain hadn't been considered a clear priority. The art of modelling used to

be an example of an approach that had been exclusively preoccupied by its own topology and coherence.

The structuring activity has obviously brought an increase of clarity regarding the modelling and its subsequent demarche, that is the implementation.

The structuring paradigm showed its weaknesses in relation to high-complexity modelling problems. Thus, we have to put together some other ideas apart from the structuring paradigm in order to be successful in such cases. We have to grant special attention to different abstractization modalities, as abstractization imposes itself as a fundamental tool for taking over the complexity. As a consequence, specialists have theorized and elaborated progressive abstractization procedures, both in relation to the universe of the data and to that of the operations. Abstract data types can be considered a kind of a synthesis of the approaches that held at their core the methodic abstractization. This is a synthesis that has anticipated the revolutionary paradigm that is represented by object orientation. Is object orientation a paradigm without problems? Aspect oriented programming suggests that the answer is no. The same message comes, although not as a reproach, from the component oriented programming, as well. Where do the most powerful criticisms of the object oriented programming come from? Surprisingly, they come from those that intensely and methodically plead for the importance of the object oriented modelling activity. Many of the object orientation concepts and principles are accompanied by indications and contraindications. Even the programmers that discerned the importance of a quality design for the success of an IT project feel that something is not in order with object orientation. It is true, no human creation can avoid the criticism of the human being himself. Even if nothing can be objected in relation to the technical aspects of a technology, which is unlikely, the problem of taste remains open. This is a problem regarding the correspondence between the technology's offerings and the user's expectancies, considering the three tiers: syntactic, semantic and pragmatic. In OOP (Objected Oriented Programming), we can state that the world can be assimilated to a structured collection of objects. Thus, we speak about a rigorously-defined set of objects that interact. Any interaction involves the existence of some actors and a communication protocol. Here lies the problem: the communication protocol is always different for every new software system. It is clear that a generally valid communication protocol cannot be specified for all the software systems. Nevertheless, we are convinced that we can "draw" a communication framework that is bound to help at specifying the communication protocols amongst the actors of a particular software system.

4 Short Inventory of Essential Problems, which are Frequent in the Modelling Activity

The systems that belong to the surrounding world are structured entities collections, due to the fact that they have the potential to operate and co-operate.

We have been aware that things are like this for a long time now. Even if we cannot precisely anticipate the future states of the systems that determine the operation of other systems, we are at least able to know, with approximation, the cause of the changes that we witness. Every systems developer, but also every researcher that studies a system, looks for the most reliable way to increase the quality of their own demarche. Consequently, what are the main issues that impact on the general endeavour for progress in relation to the systems development and modelling? Following, they are presented according to a sequence that does not yet suggest anything:

- The limited human capacity to analyse the structure and the behaviour of the systems. The limitation that we refer to in this case is quantitatively and qualitatively measurable.

We cannot think considering rhythms that resemble to the power of the continuum. We think according to rhythms that are consistent with the discrete structures, on which is founded the difficult-to-define human intelligence. The discrete systems are not necessary examples of approximate knowledge, which are prone to fast moral wear. Humans have created systems, be them theoretical or practical that have successfully passed the test of time, which is the surest proof of their quality.

- The intrinsic complexity of any creative demarche. Modelling the creativity is a priori a chimera. Nevertheless, it is within human reach to stimulate the creativity through the automation of all the activities, which together with methodical repetition and optimization become algorithmizable. Freed up from the concern to consume time for performing some routine activities, the human being has all the time in the world to be creative, that is to push forward his fight with the unknown from within himself and from the universe.
- Being characterized by the limited capacity to analyse the structure and the behaviour of the systems, and considering that any creative step features a non-conventional complexity, the human being pushes his boundaries through the use of the paradigms. Any paradigm, regardless of the field of knowledge, is an instrument for treasuring some qualitative accumulations that relate to the human creativity. Any paradigm has the role of keeping active the flame of knowledge discovery, provided it becomes a tool that is generally accessible. Thus, it can be stated that paradigms resemble to their creators: they are being born, they mature, they flourish for a while, and then they are replaced by some other paradigms.
- The most precise history of a certain field of knowledge is the story of the never-ending chain of paradigms, which have fuelled the illusion of forward movement in that respective field of knowledge. The history of the happenings inside each paradigm is the only one that has the power to legitimize the joy of living timelessness, but also the need to periodically return to its ever-renewed sources. We dare to assert that the perpetual precariousness of the paradigms, to which the human being relates as a knowledge agent, is a trick that is used with the certainty that the human reasoning is able to push the boundaries of knowledge deeper into the unknown. The goal of each paradigm, not always clearly stated, is to bring an increase of knowledge in a certain field. The initial universality of each paradigm is abandoned as soon as the achievements that have been registered in relation to it proved its conjunctural capabilities.
- Although modelling is an activity that is made possible only from inside a paradigm, which impose a certain rigour to the approach, obeying the exigencies of the rigour is, generally speaking, problematic. Humans prefer to model as if they were speaking in the mother tongue, by making slalom through ambiguities, correctness, metaphors, formal abstractions with limited reach, and some other methods that are used to explore the incomprehensible. This natural predisposition of the humans conflicts with the need to discipline the modelling activity, which is claimed by the tools producers and also by the management of the IT projects that is intensely focused towards the success of these projects.

We have presented through the problems that have been introduced, part of the reasons that provoke the seemingly chaos that thrones in the world of the modelling paradigms. This chaos is, considering a more attentive analysis, the expression of the continuous search effort that aims to enrich a given paradigm or propose a new paradigm. In this paper, we present the main ideas of a modelling paradigm that establishes the importance of the concept of interface as a methodical bridge, which links the structure and the behaviour of a software system.

5 Concepts and Fundamental Principles that Are Used in the Interface Oriented Modelling of the Software Systems

5.1 Preliminaries

In accord with the objectives of this paper, in this section we attempt to sketch the essential conceptual infrastructure, which is necessary in order to practice the interface oriented modelling in the IT industry. This infrastructure obviously refers to the concepts with the help of which the interface oriented modelling may become operational, with promising results. Figure 1 presents the role that is played by the concept of interface in the modelling activity, considering very high levels of abstraction.

In *Figure 1*, the two actors that are designated by Developer and Client abstract the modelling expert that is perceived in one of the two fundamental hypostases: models producer or models user in his own modelling activity. In any of the two hypostases, the modelling expert can address the complexity-related issues by granting the proper attention to the concept of interface. Although what Figure 1 suggests is as correct and obvious as it can be, this does not mean that any modelling expert automatically becomes a winner, if he understands the message transmitted by Figure 1. What is, in fact, the message?

The mind of the one that models has to get used to perceive systems, that correctly and coherently relate to their environment through the concept of interface. Thus, we solve two problems that have a major impact on the quality of the models that we elaborate: the streamlining of the service traffic between software systems and the assurance of the modular continuity in relation to the software systems.

We could end the presentation of this paper here but, it wouldn't be a wise decision. Considering the *Figure 1*, we cannot deduce with sufficient clarity how the interface can help to optimize the relation between the structure of a system and the behaviour that is built on this structure. In order to provide additional clarity to our demarche, we make some preliminary remarks:

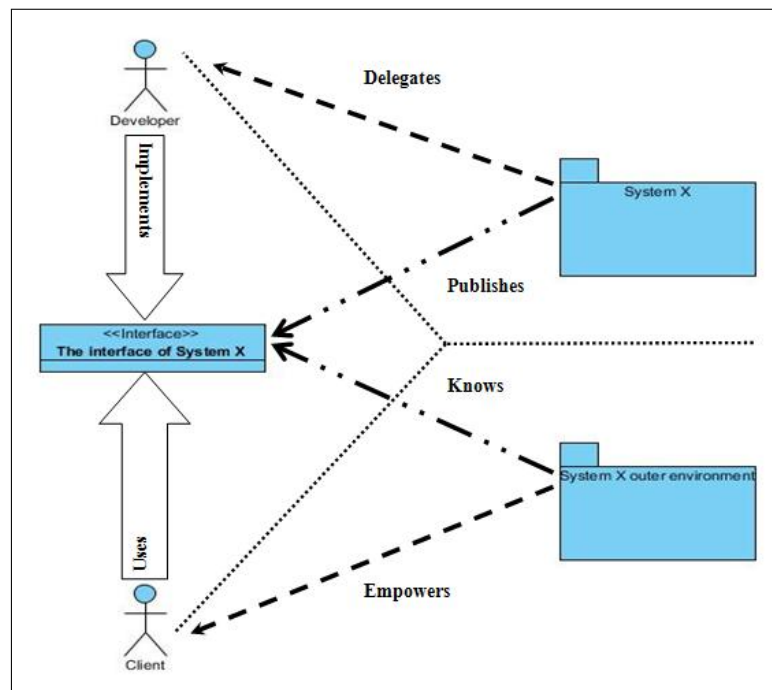


Figure 1: The interface: a fundamental tool for a modeling activity of the real world in IT

1. The modelling activity is an endeavour in which the abstractization is the essential tool that allows for the complexity to be mastered and modelled.
2. The modelling activity is spatially organized: on the horizontal, the semantics that correspond to different perspectives are structured; on the vertical, the abstraction levels of these perspectives are structured.

Consequently, it is necessary to specify the concepts, with which we operate inside the perspectives at various abstraction levels. We plan to deal with this problem, at least considering several iterations. At each iteration, we present a diagram, which is connected to an abstraction level that summarizes the concepts that exist at that level.

The beginning of the scientific knowledge passes through the concept of system. Considering that the modelling is a remarkable kind of knowledge, we can infer that the legitimate origins of the modelling belong to the area that is determined by the concept of system. The concept of system represents equally a modelling tool and, also, a tool for representing a certain type of modelling approach. As a modelling tool, the system can be explained through making use of three perspectives: the user perspective (UP), the structural perspective (SP), and the behavioural perspective (BP).

The necessity to utilize the three perspectives is relative. Representing different levels of maturation of a modelling demarche, and approached in an iterative manner, the three perspectives feature as a terminal point the decryption of the systems' behaviour, as it can be noticed in *Figure 2*. It is also natural to reflect on the links that are established between the three perspectives.

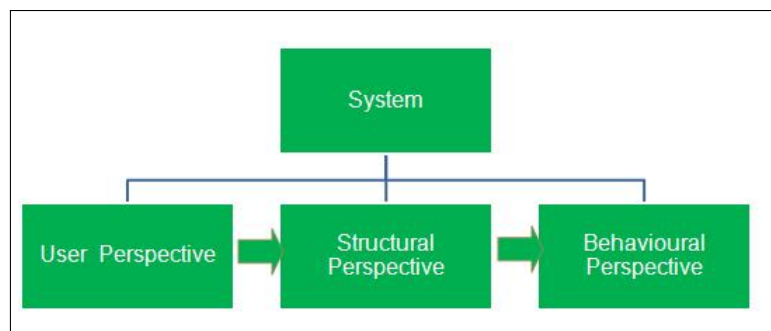


Figure 2: The main perspectives that are used for modelling a system

It can be assumed that based on the knowledge acquisition that is organized according to UP, the elaboration of SP becomes more natural, while considering an approach that favours the correct specification of BP, considering any level of detail.

Therefore, we speak about naturalness in the specification and preparation process of an optimal framework that is suitable to specify the behaviour. The link that is established between the three perspectives goes even further. This will be better understood as soon as the concepts that are part of the interface oriented abstractization base infrastructure are introduced.

5.2 The Concept of Interface in the Software Systems Modelling

Before proceeding to the actual brief description of each perspective and to their methodology driven assembly, we define, among other things, the concept of interface, while considering the highest level of abstractization.

In the practice of software systems modelling, it is called interface a protocol that allows for the communication between two objects to occur. The communication between two objects

takes place through a messages exchange. In order for the communication to be feasible, it is necessary that the sending object (the object that requires a service) knows the interface of the object that receives the message (the object that will provide the service).

This is the way the concept of interface can be understood, provided it is documented from an operational perspective. From a conceptual perspective, the notion of interface bears a series of meanings, which are probably more important for the quality of the software systems modelling activity.

An interface is a model that abstracts the messages exchange between two objects, with two essential goals:

1. The safeguarding of the fact that the objects that send messages to the object that owns the interface use a calling scheme, which remains unchanged in a reasonable time span.
2. The exemption of the object that receives messages from the task to adapt to possible changes from the environment it is related to.

In other words and considering the perspective that is induced by these definitions, an interface is correctly specified if the object it serves (service provider for the environment it is a part of) is, at the same time:

- temporarily degrevated of the responsibility to adapt to the environmental changes;
- free to modify the implementation of the services that are published through the interface.

We are now entitled to ascertain that the interface is a double-sided abstraction tool:

- Provides stability for the clients that call it, which entitle us to state that the interface features genuine structural ambitions, because it regulates, in the long run, the manner according to which the clients of the owner-object may call the object's services.
- Relieves the object that receives a messages of the task to adapt to the environmental changes; thus, the interface offers the ideal framework for the respective object's behaviour optimization.

Without trying to state that this is the optimal variant, we promote the idea that the correct transition from a system's structure to its behaviour involves the methodical utilization of the concept of interface.

With the wish to clarify the framework related to which the contents and the connections of the three perspectives in *Figure 2* are presented, we introduce in the following paragraphs, two principles that feature a great power of generalization and up-to-datedness for the general knowledge process and, consequently, for the modelling activity.

5.3 The Principle of Consistency and Deductive Conditioning of the Perspectives

The logic that underlies the modelling activity is the one that governs the general knowledge. It is known that in order to model or explain the reality, the human being needs to analyse this reality. In the case of high-complexity systems, the analysis activity is organized in such a way that the human capacity to reason sequentially may found an approach, whose outcomes are comparable to a hypothetical knowledge/modelling exercise that is based on parallel processing capabilities. It could be stated that one of the secret aspirations of the human mind is the wish

to relate to this reality not as to a sum of parts that are put together by a structure with a certain granularity, but as to a whole whose parts are articulated according to rules that are known in detail. The latter set of rules features an essential importance, which is nevertheless secondary as compared to the importance of the whole.

It is essential to provide two assertions:

The regular human perceives sets of components that are structured according to the adequate observation and research possibilities;

The demiurge perceives systems inside of which the components move according to the implacable logic of the principles that form the foundation of the universe.

These are the two extreme approaches, which can be tried in the activity of knowledge acquiring/modelling. Given the precariousness of the paradigms that are elaborated in order to support demiurge-like approaches, we can only optimize the regular human-like approach. In such an approach, the complexity is taken over in a progressive manner, while the knowledge effort is mapped on several perspectives, in order to enhance the accuracy of the analysis. Each perspective has the goal to relate with specific means to the same semantics.

The consistency of each perspective and the deductive conditioning of the perspectives are two requirements that have the value of a principle in the activity of modularization/knowledge acquiring, in general.

The hypothetical continuity of a modelling approach is assured through an accurate relation to the principle of consistency and deductive conditioning of perspectives. The problem regarding the hypothetical continuity of a modelling approach is linked to another invariant of any modelling approach: the change as an immanent attribute to any real system.

Additionally, whether it is about the law of the entropy, or about the on-the-fly re-definition of the structural equilibria, the change is an important variable in relation to any system's modelling.

Consequently, the problem can be formulated as follows: is the model that has been elaborated capable of reacting to changes according to the level at which changes operate? Or, the changes provoke shock waves that affect the system in an uncontrolled manner? The continuity of a modelling demarche would ensure that the developer enjoys some added comfort, in the case he confronts with various types of change exhibitions.

5.4 The Principle of the Perspectives Overlap in the Modelling Activity

Although it can be considered a truism, we indicate another fact that is used by abstractization techniques in the software industry, without trying to problematize: the overlap of the perspectives. The idea of the perspectives' overlap can be presented as per the following paragraphs.

Considering that an abstractization method involves a multi-perspective approach, the connection between the perspectives is not only a deductive constraint (with the meaning stated in Figure 2, which indicates the qualitative progress of the abstractization), but also the overlap inside each perspective, which confirms the quantitative interdependence of the perspectives.

The modelling represents, at each moment, a mixture of evolution in time and specific state. As an example, the elaboration of the user perspective is premise for the structural perspective, but the ingredients it is composed of are of the type UP, SP and BP. Naturally, this composite is always characterized by the name of the ingredient that is documented at the highest degree, mainly UP in this context. Thus, the iterative and incremental nature of the human approach to modelling becomes manifest. This is valid both from a natural perspective, and considering the human struggle towards attaining efficiency in relation to the modelling activities.

Following, we describe the manner according to which the concept of interface, when used

at different abstraction levels of a software system solution, can contribute to disciplining and making more efficient the specification of the perspectives, which are mentioned in Figure 2.

5.5 The Methodological Unification of the Perspectives in the Modelling Activity

Although it seems exaggerate to consider such aspects, it is necessary to note that the identification of a software system's structure is just a premise for understanding its behaviour. Possibly, an explanation for this reality resides in the different ways that are used to perceive the structure and the behaviour.

While the structure needs to be mainly invariant and flexible as an additional feature, the behaviour requires mainly flexibility and has to be invariant as an additional feature.

The above assertion should become even clearer if we added the following explanatory note: while the invariance of a system is perceived based on the rigour of the relations that are established between its component parts, the system's behaviour is perceived based on the results that it produces while it operates.

The apparent pressure that exists between the two types of perspectives (SP and BP) can be "attenuated" by carefully handling the concept of interface. The careful handling involves the correct understanding of the roles that are played by the interface in relation to the partnerships it is involved in. Thus, in the context that is determined by SP, the shape of the contract between interfaces and the implementation-related resources can modify, while the content of the contract should be invariant. This becomes clear if we watch the semantics that is presented in *Figure 3*. In the context that is determined by BP, the shape of the interactions is temporarily frozen, while the content of the interactions can suffer modifications.

Let us study the semantics that is shown in *Figure 4* in order to understand the meaning of the above statements.

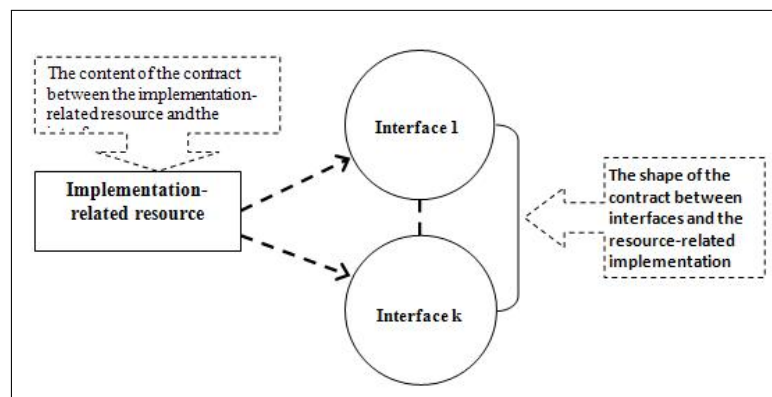


Figure 3: The role of the interfaces in the context of SP

Considering the previous considerations, we can also study the representation in *Figure 5*. This discusses the natural link between two abstractization circuits of a software system's solution:

- The **small circuit** (interior), which enhances the principle of the deductive conditioning of the perspectives, but also the principle of the perspectives overlap considering each step of the modelling effort.
- The **big circuit** (exterior), which presents the defining elements that support the consistency of the three perspectives (UP, SP, BP).

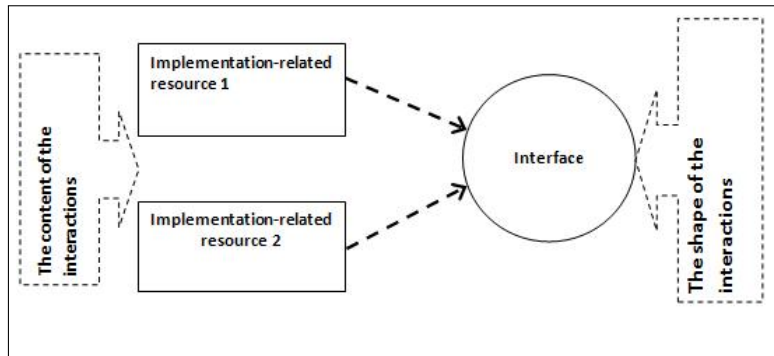


Figure 4: The role of the interfaces in the context that is determined by BP

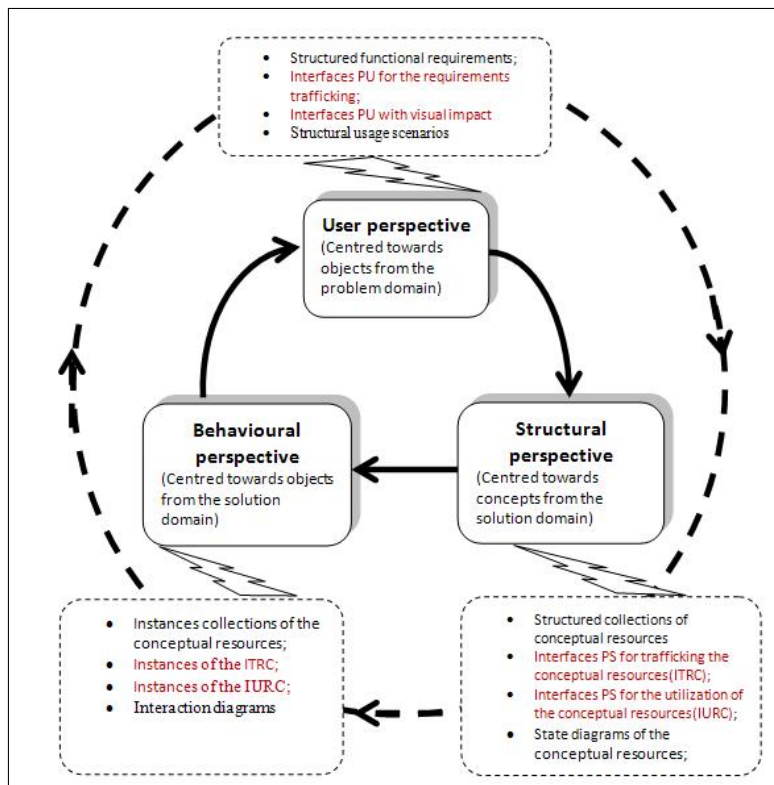


Figure 5: Unifying the perspectives in the IT systems modelling activity

6 Conclusions and Future Developments

We are now able to state that the methodological unification of the three perspectives in the modelling activity can be realized in an artisanal manner, or through methodically using the concept of interface. This involves the usage of the concept of interface, considering the particular usage perspective, as a stabilizing tool or as a flexibility-related tool in connection to different aspects of the solution, as it has been suggested by the proper remarks that have been made.

In a future paper, we plan to approach the problem of drawing a detailed framework that is bound to foster the derivation of a software system's behaviour from its structure.

Bibliography

- [1] Bloch, J.; *Effective Java - Second Edition*, Addison-Wesley, 2008.
- [2] Craig, I.; *Object-Oriented Programming Languages: Interpretation*, Springer, 2007.
- [3] Garzas, J.; Piattini, M.; *Object Oriented Design Knowledge - Principles, Heuristics and Best Practices*, Idea Group Publishing, 2007.
- [4] Martin, R.C.; *UML for Java Programmers*, Prentice Hall, 2003.
- [5] Gomez, J.; Cachero, C.; OO-H Method - Extending UML to Model Web Interfaces, *Information Modeling for Internet Applications*, 144-173, 2003.
- [6] Rector, A.; Axioms and templates: distinctions and transformations amongst ontologies, frames, and information models, *Proc. of the K-CAP Conference*, 73-80, 2013.
- [7] Espinoza, L.; Espinoza, H.; Feng, W.; Modeling a Facilities Management and Information System by UML, *Proc. of the ITNG Conference*, 65-70, 2013.
- [8] Filip, F.G.; A Decision-Making Perspective for Designing and Building Information Systems, *INT J COMPUT COMMUN*, ISSN 1841-9836, 7(2):264-272, 2012.
- [9] Marian, Z.; Czibula, G.; Czibula, I.G.; Using Software Metrics for Automatic Software Design Improvement, *Studies in Informatics and Control*, ISSN 1220-1766, 21 (3):249-258, 2012.

Estimation of the Text Skew in the Old Printed Documents

D. Brodić, Č.A. Maluckov, L. Peng

Darko Brodić, Čedomir A. Maluckov

Technical Faculty in Bor,
University of Belgrade
V. J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs, cmaluckov@tf.bor.ac.rs

Liangrui Peng

Department of Electronic Engineering,
Tsinghua University
Beijing 100084, P.R. China
penglr@tsinghua.edu.cn

Abstract: Old printed documents represent the significant part of our heritage. In order to preserve them, the digitalization is indispensable. The paper proposed a robust skew estimation method for old printed document. It is based on the connected components made by filled convex hulls around text element. The connected components are enlarged by oriented morphological operation. Then, the longest connected component is extracted. The global orientation of the document is detected by its orientation. Accordingly, document image was globally de-skewed. The algorithm is tested on synthetic and real datasets. Obtained results proved the algorithms correctness.

Keywords: document image analysis, moment methods, optical character recognition, skew adjustment.

1 Introduction

Old documents represent the part of great cultural and scientific importance. Due to age, it is quite common for such documents to suffer from degradation. Examples of degradations include shadows and variable background intensity, smudges, ink seeping, smear and strains. These degradations make image preprocessing particularly difficult and produce recognition errors. In document automatic recognition systems, the quality of the input image is crucial to final performance. There are a variety of interfering effects such as noise and skewing that appear during the scanning process. These components disturb the proceeding and decrease the performance of the recognizer. Skew correction plays an important role in the image preprocessing. A small inclination in document image can interfere in the layout analysis and consequently in the rest of the process. That's why, the identification of the object skew in the image is one of the most important tasks in digital image processing and document image analysis. It is so due to optical character recognition (OCR) system sensitivity to any skew appearance in the text.

In this paper, we deal with old printed documents like letters, technical notes, etc. They are characterized with the shape regularity as any other printed text, [1] which contain letters with similar sizes. The distance between text lines is adequate, which facilitates separation of text lines. The orientation of the text lines is similar. That considers pretty the same skew, which represents the global text skew.

A large amount of techniques has been developed in order to identify text skew. They are classified as: [1] projection profiles methods, k-nearest neighbor clustering methods, Hough transforms methods, Fourier transformation methods, cross-correlation methods, and other methods. Many of these methods have strong points as well as weaknesses. Projection profile method

is a straightforward method, which is suitable for text with uniform skew only. [2] K-nearest neighbor clustering method cannot handle incorporation of noisy subparts in text, which leads to reduced accuracy. [3] The Hough transforms method needs preprocessing stage, which defines candidate mapping points. [4] The method is complex and computer time intensive. The Fourier transforms method is even more complex. [5] The cross-correlation method is limited only to small skew angles up to 10° . [6] Interesting extension of those methods represent the incorporation of log-polar transformation. [7] However, sometimes it is unstable in application. The techniques classified as other methods are based mostly on combination techniques. They have been reputed as the most efficient ones. However, they are multistage and computer time intensive. Such methods are proposed in [8]- [9]. Preprocessing of document image is made by complex decision making. It is performed with complex geometrical filtering. The text skew is identified with the cross-correlation method applied to remain connected components. At the end, local text skew is calculated with the least square method. This technique performs local skew estimation and reliable text localization without restriction of the skew angle value.

The main contribution of this paper is the algorithm suitable for the recognition of the text skew in the old printed documents characterized with dominant skew.

Organization of the paper is as follows. Section 2 describes the algorithm. Section 3 defines text experiments. Section 4 gives the results and discusses them. Section 5 makes conclusions.

2 Proposed Algorithm

The proposed algorithm identifies the skew, which represents the dominant skew of the whole printed document. It consists of the steps that follows: 1) Uneven illumination reduction with binarization, 2) Convex hulls extraction, 3) Joining text objects with oriented binary morphology, 4) Extraction of the longest object, 5) Skew estimation of the longest object by the moments, and 6) Global de-skewing of the original document.

2.1 Uneven illumination reduction with binarization

The binarization method adopts both global and local adaptive thresholds. [10]- [11] First, multiple candidate thresholds are computed via histogram of the original gray image. Those pixels which are definitely background and foreground pixels are recorded, and the remained gray pixels will be binarized by the adaptive thresholding method. Second, we get the binarized results of the remained gray pixels by multiple candidate thresholds respectively, the statistical parameters such as run-length are calculated for each binarized images. Finally, the optimal threshold is selected when the statistical parameters are stable. If the statistical parameter analysis fails, a global threshold will be calculated by histogram analysis. After the binarization, document image is transformed into a binary matrix \mathbf{B} featuring M rows and N columns. It has two intensity levels, i.e. $B(i,j) = \{0,1\}$. Figure 1 shows the document image before and after binarization process.

2.2 Convex hull extraction

Instead of using bounding boxes, [12] the proposed algorithm exploits the convex hulls over text. Convex hull creates a smaller region around the text compared to bounding box. Hence, the probability of touching the neighbor text fragments has been reduced because of smaller contour. Upon the extraction of convex hulls, they are filled with white pixels due to complementary image. Such a text image is given with matrix \mathbf{C} . Figure 2(a) shows convex hulls extraction.



Figure 1: Document text image: (a) Before binarization, (b) After binarization

2.3 Joining text objects with oriented binary morphology

Currently, some of filled convex hulls are joined. They create short connected components (CC). The longest of them CC_{ISR} accompanies the attribute of the orientation intention. It is called initial skew rate (ISR). CC_{ISR} is extracted by the application of the longest common subsequence (LCS). [13]

$$CC_{ISR} = \max_{i,j} \left(\bigcap_{m=1}^K CC_m \right), \tag{1}$$

where $K = 1$ is the total number of CC. CC_{ISR} is shown in Figure 2(b). Its orientation is calculated by the moments (See eq. (7) for reference).

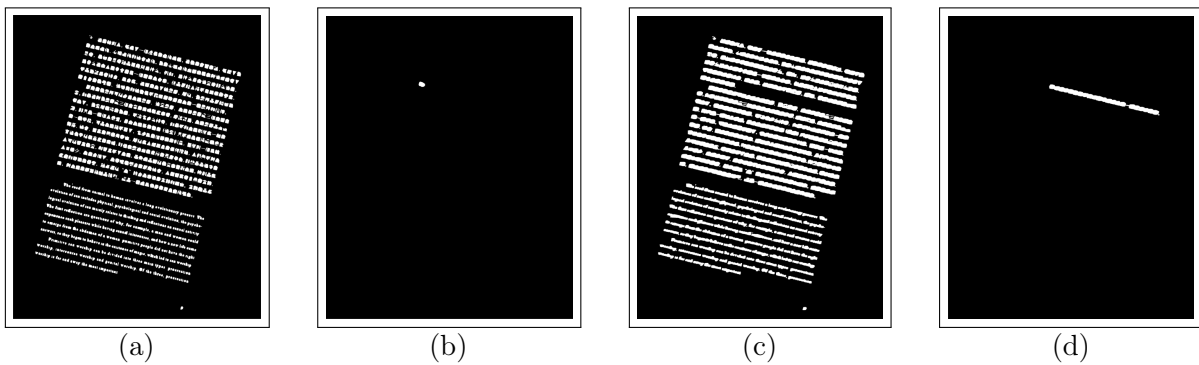


Figure 2: Document text image: (a) CC extraction (b) CC_{ISR} extraction, (c) Extended CC extraction (b) The longest CC extraction

In order to correctly estimate the text skew, connected components should be extended. Hence, morphological erosion is applied to C . This way, the adjacent CC 's are merged establishing parts of the text line. Structuring element S representing a variable width line is used. In order not to touch or join separate neighbor text lines, the width of the line should be chosen with caution. It heavily depends on each CC 's height. Empirically, it is used as 30% of the connected component's height, which means that width of structuring element S applied to each CC 's is different. Furthermore, its variability is a function of ISR, because structuring element S is skewed according to ISR orientation. Morphological operation is given as:

$$Y = C \oplus S(\angle ISR). \tag{2}$$

Figure 2(c) shows extended CC made by oriented morphology.

2.4 Extraction of the longest object

Currently, extended CCs are created. They represent partial text lines. It is clear that the longest of them CC_{LNG} incorporates the orientation which is similar to text skew. Hence, it is mandatory to extract CC_{LNG} from \mathbf{Y} . Again, it is performed with LCS method: [13]

$$CC_{LNG} = \underbrace{\max}_{i,j} \left(\bigcap_{n=1}^L CC_n \right), \quad (3)$$

where L is the total number of extended CC. CC_{LNG} is shown in Figure 2(d). Document text skew can be estimated by identifying the orientation of CC_{LNG} .

2.5 Skew estimation of the longest object by the moments

In order to estimate the skew orientation of CC_{LNG} , the moment based technique is used. Moment defines the measure of the pixel distribution in the image. It identifies global image information that depends on its contour. Moments of the binary image \mathbf{Y} featuring M rows and N columns are: [14]

$$m_{pq} = \sum_{i=1}^M \sum_{j=1}^N i^p j^q, \quad (4)$$

where p and $q = 0, 1, 2, 3, \dots, r$, and r represents the order of the moment. The central moment's μ_{pq} of the binary image \mathbf{Y} can be calculated as:

$$\mu_{pq} = \sum_{i=1}^M \sum_{j=1}^N (i - \bar{x})^p (j - \bar{y})^q. \quad (5)$$

The image feature which represents the object orientation θ is obtained from the moments. It illustrates the angle between the object and the horizontal axis. It is given as: [14]

$$\theta = \frac{1}{2} \arctan \left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}} \right), \quad (6)$$

Hence, the orientation θ of the longest object CC_{LNG} estimates the global text skew.

2.6 Global de-skewing of the original document

According to the orientation of the longest object θ , the initial document image is de-skewed. Figure 3 shows the document image before and after de-skewing process.

3 Experiments

Main goal of the experiment is the evaluation of the algorithm's ability to estimate text skew. It is performed on real and synthetic datasets. In this case, synthetic dataset consist of the samples that include single-line printed text. The samples are given in the resolution of 300 dpi. They are rotated for the angle θ , from 0° to 10° by 1° and from 10° to 40° by 5° steps around x -axis in the positive direction. It is shown in Figure 4(a).



Figure 3: Document text image: (a) With skew (b) After de-skew



Figure 4: Dataset rotation: (a) Synthetic dataset, (b) Real dataset

Real dataset consists of document image samples given in the resolution of 150 and 300 dpi. They are rotated for the angle θ , from 0° to 10° by 1° and from 10° to 40° by 5° steps around x -axis. Figure 5 shows document samples of the real dataset in Latin, Serbian Cyrillic, Chinese and Greek Cyrillic (excerpt from ICDAR 2013 text skew test).

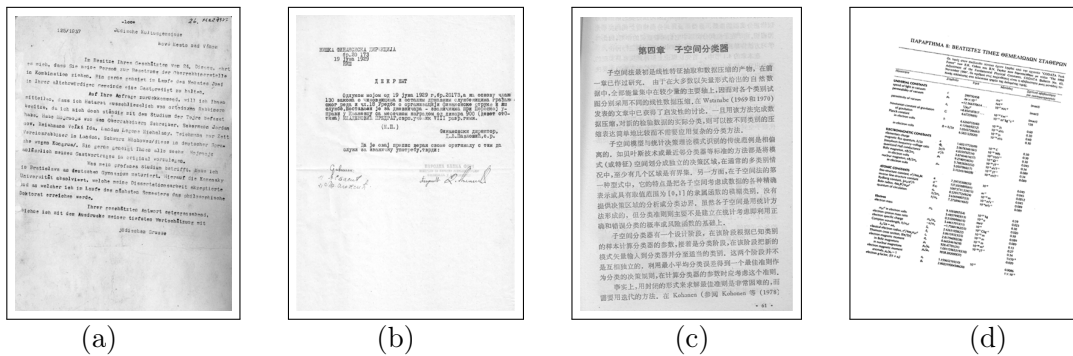


Figure 5: Samples from real dataset: (a) Latin document, (b) Serbian Cyrillic document, (c) Chinese document, (d) Greek Cyrillic document

After the algorithm’s application to dataset obtained result represents the estimated text skew. This result is compared to the reference text skew of the document samples from dataset. The evaluation of the algorithm’s result is made by the absolute deviation:

$$\Delta\theta_A = |\theta_{REF} - \theta_A|, \tag{7}$$

where θ_{REF} is the reference skew of the input text sample and θ_A is text skew estimated by the algorithm.

4 Results and Discussion

Table 1 shows the absolute deviation of global text skew for synthetic and real dataset. These result are given for the full range of rotation angles ($0^\circ - 40^\circ$).

Table 1: Absolute deviation for synthetic and real dataset

Resolution		300 dpi		150 dpi		
Dataset	Synthetic	Real	Synthetic	Real		
θ_{REF}	$\Delta\theta$	<i>isr</i>	$\Delta\theta$	$\Delta\theta$	<i>isr</i>	$\Delta\theta$
0	0.0734	1.0362	0.0376	0.0676	-0.6384	0.2490
1	0.0408	2.2297	0.0147	0.0481	0.8453	0.0329
2	0.0355	1.6194	0.1067	0.0141	1.0806	0.0005
3	0.0569	-0.0450	0.0020	0.0078	2.4496	0.2352
4	0.0001	0.7877	0.0209	0.0468	3.3863	0.0086
5	0.0188	2.2632	0.2458	0.0001	1.7279	0.2227
6	0.0118	2.9291	0.0221	0.0492	5.6370	0.1314
7	0.0265	3.9265	0.2889	0.0107	6.0385	0.0679
8	0.0740	4.9008	0.0504	0.0532	6.9414	0.0769
9	0.0604	7.0503	0.0987	0.0398	8.5425	0.0045
10	0.0470	5.5578	0.2775	0.0493	8.4272	0.9760
15	0.0586	11.9536	0.3697	0.2334	15.0123	0.0648
20	0.0680	15.9705	0.3497	0.2476	19.5475	0.1556
25	0.0793	21.9282	0.2223	0.2942	26.5648	0.1537
30	0.0797	27.0247	0.2245	0.3048	28.7789	0.2365
35	0.3266	31.9050	0.1856	0.3770	34.5256	0.2569
40	0.3553	36.8890	0.2028	0.4086	38.9246	4.6753
Average	0.0831	-	0.1600	0.1325	-	0.4440

From Table 1 results are as follows:

- for synthetic dataset given in the resolution of 300 dpi the absolute deviation is below 0.08° for the angles up to 30° and below 0.35° for the angles between 30° and 40° with the average value of 0.08° ,
- for real dataset given in the resolution of 300 dpi the absolute deviation is below 0.37° for the angles up to 40° with the average value of 0.16° ,
- for synthetic dataset given in the resolution of 150 dpi the absolute deviation is below 0.05° for the angles up to 10° and up to 0.415° for the angles between 10° and 40° with the average value of 0.13° ,
- for real dataset given in the resolution of 150 dpi the absolute deviation is below 0.97° for the angles up to 35° with the average value of 0.44° .

The proposed text skew algorithm has the average absolute deviation of 0.16° for the text skew angle θ up to 40° . Compared obtained result with the result of the algorithm without using oriented morphology, [16] the average value of absolute deviation is lower approx. 0.2° . It has quite acceptable values of the absolute deviation in the wide range of angles. Furthermore, the algorithm has been applied to different types of documents (including few examples from Document Image Skew Estimation Contest - ICDAR 2013) and different types of letters. It can be

used for documents like letters, technical articles, journals, dictionary, etc. Furthermore, above results are quite acceptable because geometrical filtering in preprocessing stage was excluded. However, proposed algorithm doesn't have the versatility of the multi-stage method proposed in [8]- [9]. This complex methods include complicated steps of geometrical filtering in preprocessing stages in order to exclude some redundant elements. However, such methods are much more computer time intensive. In further development, proposed method should be expanded with the inclusion of some additional geometrical filtering steps. This step will contribute to lower dispersion of absolute error value (up to 0.1°).

5 Conclusions

The paper proposed robust method for the estimation of global text skew. The method shows good results of global skew estimation for different resolution of test images. It is a merit of the moments exploration. Furthermore, the algorithm is suitable for text skew identification of document types like letters, technical articles, journals, dictionary, etc. Due to the exclusion of the preprocessing elements, some of redundant data were included in the process of text skew identification. Hence, further development of the algorithm should include geometrical filtering, which will lead to lower dispersion of estimated skew value.

Acknowledgment

This work was partially supported by the Grant of the Ministry of Science from Republic of Serbia, as a part of the project TR33037 and III43011 within the framework of Technological development program and the National Natural Science Foundation of China under Grant No. 61261130590.

Bibliography

- [1] Amin, A.; Wu, S. (2005); Robust Skew Detection in Mixed Text/Graphics Documents, *Proc. of 8th ICDAR*, Seoul, Korea, 247-251.
- [2] Manmatha, R.; Srimal, N. (1999); Scale Space Technique for Word Segmentation in Handwritten Manuscripts, *Proc. of 2nd ICSSTCV*, LNCS 1682, London, Great Britain, 22-33.
- [3] O'Gorman, L. (1993); The Document Spectrum for Page Layout Analysis, *IEEE Trans Pattern Anal Mach Intell*, ISSN 0162-8828, 15(11): 1162-1173.
- [4] Louloudis, G.; Gatos, B.; Pratikakis, I.; Halatsis, C. (2008); Text Line Detection in Handwritten Documents, *Pattern Recognition*, ISSN 0031-3203, 41(12): 3758-3772.
- [5] Postl, W. (1986); Detection of Linear Oblique Structures and Skew Scan in Digitized Documents, *Proc. of 8th ICPR*, Paris, France, 687-689.
- [6] Yan, H. (1993); Skew Correction of Document Images Using Interline Cross-Correlation, *CVGIP: Graphical Models and Image Processing*, ISSN 1049-9652, 55(6): 538-543.
- [7] Brodić, D.; Milivojević, Z.N. (2013); Log-polar Transformation as a Tool for Text Skew Estimation, *Elektronika Ir Elektrotehnika* ISSN 1392-1215, 19(2): 61-64.
- [8] Saragiotis, P.; Papamarkos, N. (2008); Local Skew Correction in Documents, *Int J Pattern Recognit Artif Intell* ISSN 0218-0014, 22(4): 691-710.

- [9] Makridis, M.; Nikolau, N.; Papamarkos, N. (2010); An Adaptive Technique for Global and Local Skew Correction in Color Documents, *Expert Syst Appl*, ISSN 0957-4174, 37(10): 6832-6843.
- [10] Otsu, N. (1979); A Threshold Selection Method from Gray-level Histograms, *IEEE Trans Sys, Man, Cyber*, ISSN 0018-9472, 9(1): 62-66.
- [11] Chen, Kuo-Nan; Chen, Chin-Hao; Chang, Chin-Chen (2012); Efficient Illumination Compensation Techniques for Text Images, *Digit Signal Prog* ISSN 0165-1684, 22(5): 726-733.
- [12] Brodić, D.; Milivojević, D.R. (2012); An Algorithm for the Estimation of the Initial Text Skew, *Inf Technol Control*, ISSN 1392-124X, 41(3): 211-219.
- [13] Brodić, D. (2011); The Evaluation of the Initial Skew Rate for Printed Text, *J Electr Eng*, ISSN 1335-3632, 62(3): 142-148.
- [14] Kapogiannopoulos, G.; Kalouptsidis, N. (2002); A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments, *Proc. of SPPRA*, Crete, Greece, 275-279.
- [15] Zramdini, A.; Ingold, R. (1993); Optical Font Recognition from Projection Profiles, *Electronic Publishing*, ISSN 0194-4851, 6(3): 249-260.
- [16] Brodić, D.; Milivojević, D.; Tasić, V.; Milivojević, Z. (2013); Identification of the Global Text Skew Based on the Convex Hulls, *Proc. of MIPRO*, Opatija, Croatia, 1282-1286.

SmartSteg: A New Android Based Steganography Application

D. Bucerzan, C. Ratiu, M.J. Manolescu

Dominic Bucerzan

Aurel Vlaicu University of Arad
Department of Mathematics and Computer Science
Romania, 310330 Arad, Elena Dragoi, 2
dominic@bbcomputer.ro

Crina Rațiu*

Vasile Goldis Western University of Arad
Department of Computer Science
Romania, 310414, Arad, Liviu Rebreanu, 91-93
*Corresponding author: ratiu_anina@yahoo.com

Misu-Jan Manolescu

Agora University of Oradea
Romania, 486526 Oradea, Piata Tineretului, 8
mmj@univagora.ro

Abstract: With the development of mobile devices the security issue migrates from the PC platform to this new technology. Securing confidential information on the mobile platforms has been, is and will be a topical issue for the specialists. In this paper we propose a new solution in order to provide security of digital data that is transferred through today's available platforms for communication.

We developed SmartSteg application that works on Android platform and it is able to hide and fast encrypt files using digital images of MB dimension as cover. LSB steganography is combined with a random function and symmetric key cryptography to transfer digital information, in a secure manner between smartphones that run under Android.

Keywords: Least Significant Bit (LSB), steganography, cryptography, android, SmartSteg.

1 Introduction

Communication and digital technology has changed society's daily activities, using information in all spheres of its existence, having a major economical and social impact. After rapid growth of the Internet and Mobile Networks, nowadays we witness the development of smaller, faster and high-performance mobile devices, which can support a wide range of features that were, not so long ago, the attributes of personal computers.

Mobile hand-held devices which are popularly called smart gadgets include: smart phones, tablets, e-book readers and are becoming essential to everyday social activities. These newly developed technologies make easier and cheaper the access, the processing, the storing and the transmitting of information. In this ever changing and evolving environment, establishing secure communication is an important target for researchers. Figure 1 shows briefly today's techniques widely used to secure digital information.

Cryptography and steganography are two techniques used to ensure information confidentiality, integrity and authenticity. Cryptography uses encryption to scramble the secret information in such a way that only the sender and the intended receiver are able to reveal it. Steganography hides the secret information in different carriers in such a way that it becomes difficult to detect. Commonly the carriers are media files (like images, audio, video) or other supports like communications protocols; an example is network steganography [6].

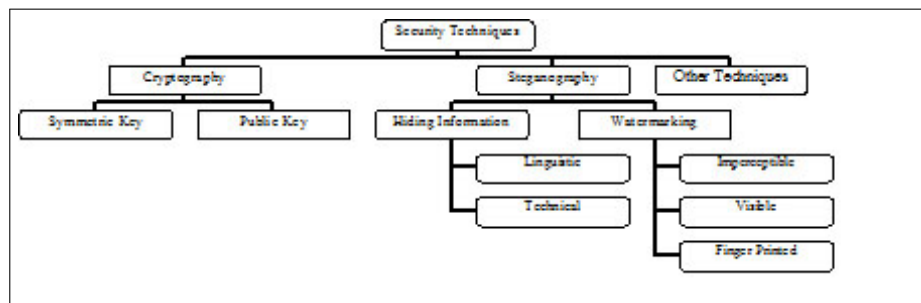


Figure 1: Security techniques

Both technologies have their limitations and this is why most of the specialists sustain that a good solution for securing the digital information is to combine the two techniques [9].

In this paper we propose a new application named SmartSteg developed to transmit secret files through Internet and Mobile Networks using a smart phone that run Android operating system. It involves technical steganography combined with symmetric key cryptography and a pseudorandom selection of the bits.

We chose to work on Android because:

- it is an open source software designed for mobile phones,
- it is well spread between mobile phones manufacturers as shown in Table 1 and Table 2,
- we have not find any reliable steganography application for Android phones that work with large files.

Company	1Q13 Units	1Q13 Market Share (%)	1Q12 Units	1Q12 Market Share (%)
Samsung	64,740.0	30.8	40,612.8	27.6
Apple	38,331.8	18.2	33,120.5	22.5
LG Electronics	10,080.4	4.8	4,961.4	3.4
Huawei Technologies	9,334.2	4.4	5,269.6	3.6
ZTE	7,883.3	3.8	4,518.9	3.1
Others	79,676.4	37.9	58,537.0	39.8

Table 1. Worldwide Smartphone Sales to End Users by Vendor in 1Q13 (Thousands of Units). Source: Gartner (May 2013) [10]

Operating System	1Q13 Units	1Q13 Market Share (%)	1Q12 Units	1Q12 Market Share (%)
Android	156,186.0	74.4	83,684.4	56.9
iOS	38,331.8	18.2	33,120.5	22.5
BlackBerry	6,218.6	3.0	9,939.3	6.8
Microsoft	5,989.2	2.9	2,722.5	1.9
Bada	1,370.8	0.7	3,843.7	2.6
Symbian	1,349.4	0.6	12,466.9	8.5
Others	600.3	0.3	1,242.9	0.8

Table 2. Worldwide Smartphone Sales to End Users by Operating System in 1Q13 (Thousands of Units). Source: Gartner (May 2013) [10]

Android may be considered a software stack for mobile devices that includes an operating system, middle ware and key applications [8]. Android architecture includes five layers, as shown in figure 2: applications layer, application framework layer, libraries, Android runtime, Linux kernel.

Linux kernel is the basis of Android architecture and it supports security, memory management, process management, network stack, and device driver model. Android's libraries are written in C/C++ and include: standard C system library, media libraries including MPEG4, H.264, MP3, JPG, and PNG, surface manager for display subsystem, LibWebCore as a web browser engine, 2D graphics engine SGL, 3D graphics libraries, FreeType for font rendering and SQLite, a lightweight relational database engine [8]. Android runtime is the engine which authorizes the applications. It includes Dalvik virtual machine and core android libraries [4]. The application framework layer includes the application manager, which allows developers to use the features of Android operating system. The application layers include all native and third party applications.

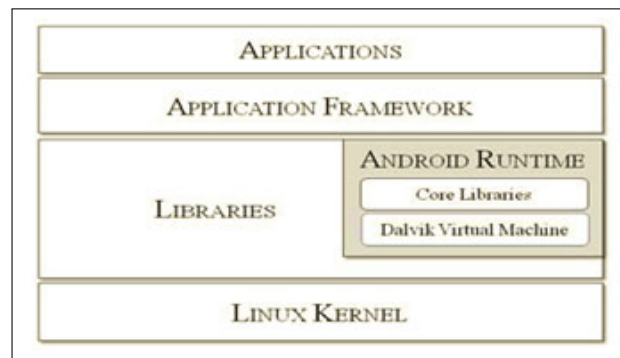


Figure 2: Android architecture [8]

2 Related Works

Image steganography has drawn attention of researchers because of its applicability in today's digitalized world. There are many applications in this direction mainly developed on computers running different operation systems.

A new approach in steganography that drawn our attention is the project developed by Mazurczyk, Karas and Szczypiorski on Skype steganography. The researchers method lies in how Skype manages silence sequences. The application encrypts the secret data and embeds it in a silence package when ones appear [3].

An interesting idea of a robust algorithm of steganography resistant to steganalytic attacks is presented in the research made by Mare [2].

Regarding Android for mobile phones, there are few reliable projects on steganography. In the following we discuss some of them.

Most of the applications that exist on Android smart phones, embeds only short sequence of characters, none of them embeds entire files. This excludes the possibility to traffic secretly images or large documents.

White and Martina developed an application Android based that uses steganography to hide a short text message in an audio message recorded by the user and then share that message [7].

MoBiSiS is an application that implements a steganographic algorithm Android based. It is able to send the image that covers the secret message via the Multimedia Messaging Service

(MMS). The cover image can be retrieved from the device's message inbox. The disadvantage of this application is the size of the cover image with the secret message embedded which must be less than 30 KB [1].

Similar applications with MoBiSiS having the same limitations, are MobiStego [12] and Pixelknot [13] both available on Google Play.

Rughani in his work [5] regarding the limitation of implementing steganography on smart phones sustains some ideas that we do not agree with:

- There cannot be a common algorithm - since there are many smart phone manufacturers. It is almost difficult to develop a common algorithm for steganography.
 - Our opinion - there are three main players (see Table 3) in the operating systems market for smart phones (Android, IO's, Windows 8) and we extend our project to all these operating systems.
- "Smart phones are small computing devices - even though smart phones are smarter than mobile phones they are not as efficient as traditional computer like desktop or laptop".
 - Our opinion - smart phones support complex, efficient and fast algorithms, fact sustained by the present work.

Operating System	1Q13 Shipment Volume	1Q13 Market Share	1Q12 Shipment Volume	1Q12 Market Share
Android	162.1	75.0%	90.3	59.1%
iOS	37.4	17.3%	35.1	23.0%
Windows Phone	7.0	3.2%	3.0	2.0%
BlackBerry OS	6.3	2.9%	9.7	6.4%
Linux	2.1	1.0%	3.6	2.4%
Symbian	1.2	0.6%	10.4	6.8%
Others	0.1	0.0%	0.6	0.4%

Table3. Top Five Smartphone Operating Systems, 1Q 2013 (Units in Millions). Source: IDC (May 2013) [11]

We also considered in our research the steganography applications available on Google Play. Most of them treat steganography and the secrecy of the communication lightly. None of them offers the advantages of SmartSteg presented in this paper.

3 The Proposed Solution

Based on Kerckhoff's principle we choose to work with secret key stegano algorithm because in this case no unauthorized person should be able to extract the secret information even the specifications of the algorithm are public.

Our proposed application, SmartSteg, works on Android smart phones. We select BMP Bitmap format for the cover images because it is a lossless format and allows embedding large quantity of information. The designed programming language for mobile application that runs Android Operating Systems is JAVA using Eclipse environment.

In our study due to Android's support of multiple devices we have encountered some problems regarding how Android manages stored images both on SD card and internal memory of a device and on different versions of Android.

Most of the applications developed in this domain are using an image view tool to obtain the cover image. The image view tool does not access directly the original image file. It makes a copy of the original image file and transforms it in an (.png) image type no matter the type of the original image. This technique reduces very much the dimension of the cover image and this is not proper for LSB because it reduces the quantity of secret information which is to be hidden. SmartSteg is able to manipulate carrier images of MB dimensions usually transferred through Internet and Mobile Networks.

This is the main reason why we worked to find a way to access the original digital image file stored on SD card or in phone's memory. This is quite a challenge on Android platform because the way to access the system root folder is different depending on Android's edition and is not widely spread among programmers.

To process the data SmartSteg follows these steps:

- Cover image, secret file, and the secret key are loaded into application.
- SmartSteg verifies the dimension of the two files (cover image and secret file) to see if they are suitable.
- The secret file, its dimension and its execution are encrypted by means of a stream cipher algorithm using the secret key. The encrypted bits are stored in a temporary array.
- LSB algorithm starts to embed secret bits inside the cover image file using the pseudo random function completed with modulo 3 operations. The purpose of this random algorithm is to spread the secret message over the cover in a rather random manner. The purpose is to create confusion for the possible steganalitic attack. Since the secret message will normally not cover the entire support file the embedding process will continue until end of cover.
- The cover image with the secret file embedded is saved and can be transmitted to an intended receiver via e-mail.

4 Design and Implementation

The characteristics of smart phones used for testing:

- Samsung GT-S5670, Android: v 2.3.4, Processor: 600 MHz
- Samsung Galaxy Nexus I9250, Android: v4.0, Processor: Dual-core 1.2 GHz Cortex-A9

The characteristics of the cover image:

- Type: BMP file
- Dimension: 4.03MB
- Pixel arrangement: 1372x1029

Types of secret files:

- txt file: 400 kb that is approximately 100 printed A4 pages
- JPG file: 402kb, 3000x1682 shown in figure 5
- different image files: BMP, JPG; archives; pdf.



Figure 3: Cover image: original and with secret message embedded



Figure 4: Secret image file embedded in the cover image

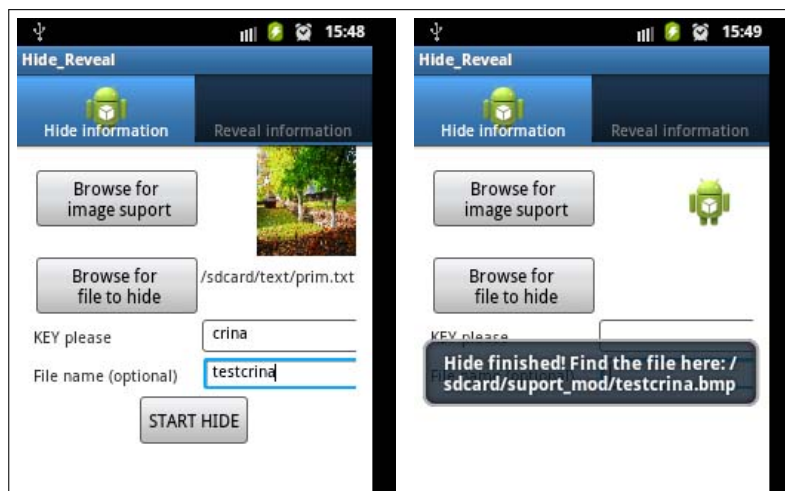


Figure 5: User interface of SmartSteg

5 Conclusions and Future Works

In our SmartSteg application, we used a sort of LSB steganography on BMP files. The proposed algorithm has reached a very good processing speed. This is significant result, considering that we manage files of MB dimension on smart phones. So far, the performance of mobile phones have exceeded our expectations concerning the execution time necessary to encode, to hide, to decode or reveal secret information even using files of MB size. The dimension of the carrier file must be approximately eight times larger than the one of the secret file.

Advantages of the proposed model:

- Combines LSB steganography with stream cipher cryptography and random algorithm on smart phone devices.
- Works with large files both the carrier BMP file and the secret file.
- Hides a large variety of files.
- Execution time practical immediately.
- The security of the proposed model lies in the usage of secret key.
- Works on different versions of Android operating system.
- The sender can take pictures with phone camera and then converts them into BMP files using them as image cover.

Future work in our research is to develop a new version of SmartSteg that is windows based. The purpose of this idea is to permit secret communication between computes and smart phones.

Bibliography

- [1] I. Rosziati, L. C. Kee, MoBiSiS: An Android-based Application for Sending Stego Image through MMS, *ICCGI 2012 : The Seventh International Multi-Conference on Computing in the Global Information Technology*, 115–120, 2012.
- [2] S. F. Mare, *Advanced Steganographic algorithms and architectures*, Teze de doctorat ale UPT, Seria 10, Nr.40, Editura Politehnica, 2012.
- [3] W. Mazurczyk, M. Karas, K. Szczypiorski, SkyDe: a Skype-based Steganographic Method, *Int J Comput Commun*, ISSN 1841-9836, 8(3):432-443, June, 2013.
- [4] D. S. Purkayastha, N. Singhla, Android Optimization: A Survey, *International Journal of Computer Science and Mobile Computing - A Monthly Journal of Computer Science and Information Technology*, 2(6):46-52, June 2013.
- [5] P. H. Rughani, H. N. Pandya, Steganography on Android Based Smart Phones, *International Journal of Mobile & Adhoc Network*, 2(2):150-152, May 2012.
- [6] K. Szczypiorski, Steganography in TCP/IP Networks. State of the Art and a Proposal of a New System-HICCUPS, <http://www.tele.pw.edu.pl/krzysiek/pdf/steg-seminar-2003.pdf>, visited on 01.09.2013.

- [7] T. F. M. White, J. E. Martina, Mobile Steganography Embedder, 11 SBSEg Simposio Brasileiro Em Seguranca Da Informacao E De Sistemas Computacionais, Bsalia-DF, 6 a 11 de Novembro de 2011, <http://www.peotta.com/sbseg2011/resources/downloads/wticg/91964.pdf>, visited on 01.09.2013.
- [8] H-J. Yoon, A Study on the Performance of Android Platform, *International Journal on Computer Science and Engineering*, 4:532-537, 04 April 2012.
- [9] B. B. Zaidan, A. A. Zaidan, A. K. Al-Frajat and H. A. Jalab, On the Defferences between Hiding Information and Cryptography techniques: An Overview, *Journal of Applied Sciences*, 10(15): 1650-1655, 2010.
- [10] ***, Gartner Press Release, *Gartner Says Asia, Pacific Led Worldwide Mobile Phone Sales to Growth in First Quarter of 2013*, 14.05.2013, <http://www.gartner.com/newsroom/id/2482816>, visited on 01.09.2013.
- [11] ***, IDC Press Release, Android and iOS Combine for 92.3% of All Smartphone Operating System Shipments in the First Quarter While Windows Phone Leapfrogs BlackBerry, According to IDC, 16.05.2013, <http://www.idc.com/getdoc.jsp?containerId=prUS24108913>, visited on 01.09.2013.
- [12] MobiStego: <http://play.google.com/store/apps/details?id=it.mobistego>, visited on 01.09.2013.
- [13] Pixelknot: <http://guardianproject.info/apps/pixelkont/>, visited on 01.09.2013.

Asymptotically Unbiased Estimator of the Informational Energy with kNN

A. Cațaron, R. Andonie, Y. Chueh

Angel Cațaron

Electronics and Computers Department
Transylvania University of Brașov, Romania
cataron@unitbv.ro

Răzvan Andonie*

1. Computer Science Department
Central Washington University, Ellensburg, USA
andonie@cwu.edu
2. Electronics and Computers Department
Transylvania University of Brașov, Romania
*Corresponding author

Yvonne Chueh

Department of Mathematics
Central Washington University, Ellensburg, USA
chueh@cwu.edu

Abstract: Motivated by machine learning applications (e.g., classification, function approximation, feature extraction), in previous work, we have introduced a non-parametric estimator of Onicescu's informational energy. Our method was based on the k -th nearest neighbor distances between the n sample points, where k is a fixed positive integer. In the present contribution, we discuss mathematical properties of this estimator. We show that our estimator is asymptotically unbiased and consistent. We provide further experimental results which illustrate the convergence of the estimator for standard distributions.

Keywords: machine learning, statistical inference, asymptotically unbiased estimator, k -th nearest neighbor, informational energy.

1 Introduction

Inference is based on a strong assumption: using a *representative* training set of samples to infer a model. In this case, we select a random sample of the population, perform a statistical analysis on this sample, and use these results as an estimation to the desired statistical characteristics of the population as a whole. The more representative the sample is, the higher our confidence level reaches so that the statistical results obtained by using this sample are indeed a good estimation to the desired population parameters. We gauge the representativeness of a sample by how well its statistical characteristics reflect the probabilistic characteristics of the entire population. Many standard techniques may be used to select a representative sample set [15]. However, if we do not use expert knowledge, selecting the most representative training set from a given dataset was proved to be computationally difficult (NP-hard) [10]. The problem is actually more difficult, since in most applications the complete dataset is unknown, or too large to be analyzed. Therefore, we have to rely on a more or less representative training set.

A critical aspect of many machine learning approaches is how well an information theory measure is estimated from the available training set. This relates to a fundamental concept in statistics: probability density estimation. *Density estimation* is the construction of an estimate of the density function from the observed data [20]. We will refer here only to *nonparametric*

estimation, where less rigid assumptions will be made about the distribution of the observed data. Although it will be assumed that the distribution has the probability density f , the data will be allowed to speak for themselves in determining the estimate of f more than would be the case if f were constrained to fall in a given parametric family. A common measure used in machine learning is mutual information (MI). Several methods were proposed for MI estimation [18], [22], [13]: histogram based estimators, kernel density estimators, B-spline estimators, k -th nearest neighbor (kNN) estimators, and wavelet density estimators.

Estimating entropy and MI is known to be a non-trivial task [4]. Naïve estimations (which attempt to construct a histogram where every point is the center of a sampling interval) are plagued with both systematic (bias) and statistical errors. An ideal estimator does not exist, instead the choice of the estimator depends on the structure of data to be analyzed. It is not possible to design an estimator that minimizes both the bias and the variance to arbitrarily small values. The existing studies have shown that there is always a delicate trade off between the two types of errors [4].

MI is generally based on the classical Shannon type MI. However, it is computationally attractive to use one of its generalized forms: the Rényi divergence measure, which uses Rényi's quadratic entropy. The reason is that, as proved by Principe *et al.*, Rényi's quadratic entropy (and Rényi's divergence measure) can be estimated from a set of samples using Parzen's windows approach [19]. The MI and Rényi's divergence measure are equivalent, but only in the limit $\alpha = 1$, where α is the order of Rényi's divergence measure [19].

A unilateral dependency measure can be derived from Onicescu's informational energy (IE). This measure proved to be an efficient alternative to MI, and we have estimated it from sample datasets using the Parzen windows approach. We used this approach in classification and feature weighting [2], [5], [6], [3], [7]. An important drawback of this approach is the fact that Parzen windows estimate cannot be applied on continuous spaces. This is also true for Shannon's type MI. Therefore, This means an important machine learning domain - continuous function approximation (or prediction), is left out.

In previous work [8], we introduced a kNN IE estimator which may be used to approximate the unilateral dependency measure both in the discrete and the continuous case. An important theoretical aspect was not discussed yet: the asymptotic behavior of this estimator in terms of unbiasedness and consistency. Generally, any statistic whose mathematical expectation is equal to a parameter is called *unbiased* estimator of that parameter. Otherwise, the statistic is said to be *biased*. Any statistic that converges asymptotically to a parameter is called *consistent* estimator of that parameter [12].

Consistent and unbiased are not equivalent. A simple example of a biased consistent estimator is if the mean of samples x_1, x_2, \dots, x_n is estimated by $1/n \sum x_i + 1/n$. This estimate is biased but consistent, since it approaches asymptotically the correct value. An *asymptotically unbiased* estimator is an estimator that is unbiased as the sample size tends to infinity. Some biased estimators are asymptotically unbiased but all unbiased estimators are asymptotically unbiased. The previous estimator is biased but asymptotically unbiased. One way to prove that an estimator is consistent is to prove that it is asymptotically unbiased and the variance goes to zero.

This gives the motivation for the present work. We show that our IE estimator is asymptotically unbiased and consistent. This will imply that the estimator is "good".

First, we summarize (Section 2) the IE and the kNN method. Section 3 describes our IE approximation method, including the novel theoretical results. After the experimental results, exposed in Section 4, we conclude with final remarks and a description of future work (Section 5).

2 Background

2.1 Onicescu's Informational Energy

Generally, information measures refer to uncertainty. Since Shannon defined his probabilistic information measure in 1948, many other authors, with Rényi, Daroczy, Bongard, Arimoto, and Guiaşu among them, have introduced new measures of information. However, information measures can also refer to certainty, and probability can be considered as a measure of certainty. More general, any monotonically growing and continuous probability function can be considered as a measure of certainty. For instance, Onicescu's IE was interpreted by several authors as a measure of expected commonness, a measure of average certainty, or as a measure of concentration.

For a continuous random variable X with probability density function $f(x)$, the IE is [11, 17]:

$$IE(X) = \int_{-\infty}^{+\infty} f^2(x) dx \quad (1)$$

2.2 The nearest neighbor method

Although classification remains the primary application of kNN, we can use it to do density estimation also. Since kNN is non parametric, it can do estimation for arbitrary distributions. The idea is very similar to use of Parzen window. Instead of using hypercube and kernel functions, here we do the estimation as follows.

The kNN estimators represent an attempt to adapt the amount of smoothing to the "local" density of data. The degree of smoothing is controlled by an integer k , chosen to be considerably smaller than the sample size; typically $k \approx n^{1/2}$. Define the distance $d(x, y)$ between two points on the line to be $|x - y|$ in the usual way, and for each t define $d_1(t) \leq d_2(t) \leq \dots \leq d_n(t)$ to be the distances, arranged in ascending order, from t to the points of the sample.

The kNN density estimate $f(t)$ is defined by [20]:

$$\hat{f}(t) = \frac{k}{2nd_k(t)} \quad (2)$$

The kNN was used for non-parametrical estimate of the entropy based on the k -th nearest neighbor distance between n points in a sample, where k is a fixed parameter and $k \leq n - 1$. Based on the first nearest neighbor distances, Leonenko *et al.* [14] introduced an asymptotic unbiased and consistent estimator H_n of the entropy $H(f)$ in a multidimensional space. When the sample points are very close one to each other, small fluctuations in their distances produce high fluctuations of H_n . In order to overcome this problem, Singh *et al.* [21] defined an entropy estimator based on the k -th nearest neighbor distances. A kNN estimate of the Kullback-Leibler divergence was obtained by Wang *et al.* in [23]. A mean of several kNN estimators corresponding to different values of k was used by Faivishevsky *et al.* in [9] for developing a smooth estimator of differential entropy, mutual information, and divergence.

According to [22], kNN MI estimation outperforms histogram methods. kNN works well if the value of k is optimally chosen. However, there is no model selection method for determining the number of nearest neighbors k . This is a limitation of the kNN estimation.

3 Estimation of the Informational Energy

We are ready now to introduce our kNN method for IE approximation, using results from our previous work [8]. The described theoretical properties are however novel. Mathematical proofs are omitted, since they would not fit into the page limit of this paper.

The IE can be easily computed if the data sample is extracted from known distributions. When the underlying distribution of data sample is unknown, the IE has to be estimated. More formally, our goal is to estimate (1) from a random sample X_1, X_2, \dots, X_n of n d -dimensional observations from a distribution with the unknown probability density $f(x)$. This problem is even more difficult if the number of available points is small.

The $IE_{empirical}$ is not a good estimate especially when the relative frequencies are far from the true probabilities. This is generally the case for small datasets and, in accordance to the central limit theorem, for an increasing number of samples, $IE_{empirical}$ converges probabilistically to IE .

The IE is the average of $f(x)$, therefore we have to estimate $f(x)$. The n observations from our samples have the same probability $\frac{1}{n}$. A convenient estimator of the IE is:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^n \hat{f}(X_i). \tag{3}$$

We will determine first the probability density $P_{ik}(\epsilon)$ of the random distance $R_{i,k,n}$ between a fixed point X_i and its k -th nearest neighbor from the remaining $n - 1$ points. Probability $P_{ik}(\epsilon)d\epsilon$ of the k -th nearest neighbor to be within distance $R_{i,k,n} \in [\epsilon, \epsilon + d\epsilon]$ from X_i , $k - 1$ points at a smaller distance and $n - k - 1$ at a larger distance can be expressed in terms of the trinomial formula [9]:

$$P_{ik}(\epsilon)d\epsilon = \frac{(n - 1)!}{1!(k - 1)!(n - k - 1)!} dp_i(\epsilon) p_i^{k-1} (1 - p_i)^{n-k-1},$$

where $p_i(\epsilon) = \int_{\|x - X_i\| < \epsilon} f(x)dx$ is the mass of the ϵ -ball centered at X_i and $\int P_{ik}(\epsilon)d\epsilon = 1$.

We can express the expected value of the $p_i(\epsilon)$ using the probability mass function of the trinomial distribution:

$$\begin{aligned} E_{P_{ik}(\epsilon)}(p_i(\epsilon)) &= \int_0^\infty P_{ik}(\epsilon) p_i(\epsilon) d\epsilon = \\ &= k \binom{n - 1}{k} \int_0^1 p^{k-1} (1 - p)^{n-k-1} p dp = \\ &= k \binom{n - 1}{k} \int_0^1 p^{(k+1)-1} (1 - p)^{(n-k)-1} dp. \end{aligned}$$

This equality can be reformulated using the *Beta* function:

$$B(m, n) = \int_0^1 x^{m-1} (1 - x)^{n-1} = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m + n)}.$$

We obtain:

$$\begin{aligned} E_{P_{ik}(\epsilon)}(p_i(\epsilon)) &= k \binom{n - 1}{k} \frac{\Gamma(k + 1)\Gamma(n - k)}{\Gamma(n + 1)} = \\ &= k \frac{(n - 1)!}{(n - k - 1)!k!} \frac{k!(n - k - 1)!}{n!}, \end{aligned}$$

which can be rewritten as:

$$E_{P_{ik}(\epsilon)}(p_i(\epsilon)) = \frac{k}{n}. \tag{4}$$

On the other hand, assuming that $f(x)$ is almost constant in the entire ϵ -ball around X_i [9], we have:

$$p_i(\epsilon) \approx V_1 R_{i,k,n}^d f(X_i),$$

where we denote the volume of the ball of radius $\rho_{r,n}$ in a d -dimensional space by:

$$V_{\rho_{r,n}} = V_1 \rho_{r,n}^d = \frac{\pi^{\frac{d}{2}} \rho_{r,n}^d}{\Gamma(\frac{d}{2} + 1)}.$$

V_1 is the volume of the unit ball and $R_{i,k,n}$ is the Euclidean distance between the reference point X_i and its k -th nearest neighbor. This means that $V_1 R_{i,k,n}^d$ is the volume of the d -dimensional ball of radius $R_{i,k,n}$.

We obtain the expected value of $p_i(\epsilon)$:

$$E(p_i(\epsilon)) = E(V_1 R_{i,k,n}^d f(X_i)) = V_1 R_{i,k,n}^d \hat{f}(X_i). \quad (5)$$

Equations (4) and (5) both estimate $E(p_i(\epsilon))$. Their results are approximatively equal:

$$V_1 R_{i,k,n}^d \hat{f}(X_i) = \frac{k}{n},$$

That is:

$$\hat{f}(X_i) = \frac{k}{n V_1 R_{i,k,n}^d}, i = 1 \dots n. \quad (6)$$

This is the estimate of the probability density function. By substituting $\hat{f}(X_i)$ in (3), we finally obtain the following IE approximation:

$$\hat{IE}_k^{(n)}(f) = \frac{1}{n} \sum_{i=1}^n \frac{k}{n V_1 R_{i,k,n}^d}. \quad (7)$$

We have introduced this result in [8]. Our main question now is to analyze its asymptotic behavior.

Consistency of an estimator means that as the sample size gets large, the estimate gets closer and closer to the true value of the parameter. Unbiasedness is a statement about the expected value of the sampling distribution of the estimator. The ideal situation, of course, is to have an unbiased consistent estimator. This may be very difficult to achieve.

Yet unbiasedness is not essential, and just a little bias is permitted as long as the estimator is consistent. Therefore, an asymptotically unbiased consistent estimator may be acceptable. In the following, we will use the following mathematical property (from [16]): An asymptotically unbiased estimator with asymptotic zero variance is consistent.

We are ready now to state our theoretical results:

1. The informational energy estimator $\hat{IE}_k^{(n)}(f)$ is asymptotically unbiased.
2. $\lim_{n \rightarrow \infty} Var \left[\hat{H}_k^{(n)}(f) \right] = 0$.

Therefore, we can conclude that the $\hat{IE}_k^{(n)}(f)$ estimator is consistent.

4 Experiments

When the distribution of a sample is unknown, the statistical measures cannot be calculated directly, and we have to use an estimate. The quality of an estimator can be determined by studying its asymptotic behavior. We proved that the informational energy estimator \hat{IE} is

asymptotically unbiased and consistent. It provides an approximation of the informational energy regardless of the distribution where the sample was drawn from. It is interesting to compare the estimated value of the IE with its real value. For an unidimensional distribution, we can achieve this goal by generating a random sample from a known distribution $f(x)$, with x_{min} and x_{max} being the minimum / maximum values. Then, the informational energy of the distribution $f(x)$ on the subdomain $\mathcal{D}_{sample} = \{x|x \in [x_{min}, x_{max}]\}$ is

$$IE_{\mathcal{D}_{sample}} = \int_{x_{min}}^{x_{max}} f^2(x)dx = \int_{\mathcal{D}_{sample}} f^2(x)dx, \quad (8)$$

while the estimated informational energy \widehat{IE} is given by the formula (7). The information energy of the same distribution has a fixed value when it is computed on its entire definition domain \mathcal{D} :

$$IE_{\mathcal{D}} = \int_{\mathcal{D}} f^2(x)dx. \quad (9)$$

Our experiments focus on the following distributions: Exponential, unidimensional Gaussian, Beta, Cauchy, Gamma, and Weibull. We use the R programming language and environment functions to generate the random samples from each distribution, with the parameters listed in Tables 1–6. The first line in each table contains: the probability density function of the distribution, the support of this function (which is the domain \mathcal{D}), the values of the parameters, and the IE computed with formula (9).

Sample size is the number of values from the random sample, and *Range* is the interval limited by the minimum / maximum values from the sample used to compute the IE with formula (8). In order to study the asymptotically unbiasedness and consistency of the estimator, we determine the value of \widehat{IE} for samples with 10, 100, 1000 values, and with increasing values of k , from 1 to the squared root of the sample size [20].

Table 1: Exponential distribution

$f(x) = \theta e^{-\theta x}, x \geq 0, \theta = 3, IE_{\mathcal{D}} = 1.5$						
Sample size: 10; Range: [0.022, 0.777]; $IE_{\mathcal{D}_{sample}} = 1.3$						
k	1	2	3			
\widehat{IE}	8.133	2.464	1.023			
Sample size: 100; Range: [0.0007, 2.599]; $IE_{\mathcal{D}_{sample}} = 1.493$						
k	1	2	3	5	7	9
\widehat{IE}	4.953	2.312	2.289	2.167	1.854	1.743
Sample size: 1000; Range: [0.0001, 2.237]; $IE_{\mathcal{D}_{sample}} = 1.499$						
k	1	2	3	10	20	30
\widehat{IE}	6.829	3.093	2.269	1.605	1.527	1.507

In general, the expected behavior was confirmed by experiments: the larger the sample size n , the more accurate estimation of the informational energy.

Table 2: Unidimensional Gaussian distribution

$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu = 0, \sigma = 1, IE_{\mathcal{D}} = 0.282$						
Sample size: 10, Range: [-1.018, 1.395], $IE_{\mathcal{D}_{sample}} = 0.254$						
k	1	2	3			
\hat{IE}	2.559	0.444	0.402			
Sample size: 100, Range: [-2.263, 2.484], $IE_{\mathcal{D}_{sample}} = 0.281$						
k	1	2	3	5	7	9
\hat{IE}	0.998	0.462	0.333	0.286	0.275	0.271
Sample size: 1000, Range: [-3.596, 2.781], $IE_{\mathcal{D}_{sample}} = 0.282$						
k	1	2	3	10	20	30
\hat{IE}	1.419	0.526	0.421	0.315	0.296	0.293

Table 3: Beta distribution

$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}(1-x)^{\beta-1}x^{\alpha-1}, 0 \leq x \leq 1, \alpha = 2, \beta = 3, IE_{\mathcal{D}} = 1.371$						
Sample size: 10, Range: [0.194, 0.773], $IE_{\mathcal{D}_{sample}} = 1.170$						
k	1	2	3			
\hat{IE}	4.873	2.452	2.083			
Sample size: 100, Range: [0.010, 0.842], $IE_{\mathcal{D}_{sample}} = 1.369$						
k	1	2	3	5	7	9
\hat{IE}	5.084	2.588	2.167	1.899	1.522	1.523
Sample size: 1000, Range: [0.010, 0.930], $IE_{\mathcal{D}_{sample}} = 1.371$						
k	1	2	3	10	20	30
\hat{IE}	101.969	2.617	2.103	1.516	1.450	1.430

Table 4: Cauchy distribution

$f(x) = \frac{b}{\pi[(x-m)^2+b^2]}, x \in R, m = 0, b = 1, IE_{\mathcal{D}} = 0.159$						
Sample size: 10, Range: [-29.068, 61.499], $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3			
\hat{IE}	1.342	0.253	0.193			
Sample size: 100, Range: [-19.543, 17.052], $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3	5	7	9
\hat{IE}	37.599	0.204	0.188	0.154	0.158	0.158
Sample size: 1000, Range: [-232.181, 165.562], $IE_{\mathcal{D}_{sample}} = 0.159$						
k	1	2	3	10	20	30
\hat{IE}	0.859	0.343	0.284	0.170	0.164	0.161

Table 5: Gamma distribution

$f(x) = \frac{x^{\alpha-1}e^{-\frac{x}{\theta}}}{\Gamma(\alpha)\theta^\alpha}, x \geq 0, \theta = 1, \alpha = 3, IE_{\mathcal{D}} = 0.187$						
Sample size: 10, Range: [1.381, 5.340], $IE_{\mathcal{D}_{sample}} = 0.156$						
k	1	2	3			
\hat{IE}	0.204	0.232	0.240			
Sample size: 100, Range: [0.556, 9.053], $IE_{\mathcal{D}_{sample}} = 0.186$						
k	1	2	3	5	7	9
\hat{IE}	0.624	0.318	0.285	0.264	0.233	0.232
Sample size: 1000, Range: [0.092, 11.866], $IE_{\mathcal{D}_{sample}} = 0.187$						
k	1	2	3	10	20	30
\hat{IE}	1.768	0.344	0.280	0.210	0.201	0.196

Table 6: Weibull distribution

$f(x) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha e^{(\frac{x}{\beta})^\alpha}}, x \geq 0, \alpha = 3, \beta = 4, IE_{\mathcal{D}} = 0.213$						
Sample size: 10, Range: [2.040, 5.202], $IE_{\mathcal{D}_{sample}} = 0.191$						
k	1	2	3			
\hat{IE}	1.315	0.583	0.523			
Sample size: 100, Range: [0.899, 7.295], $IE_{\mathcal{D}_{sample}} = 0.212$						
k	1	2	3	5	7	9
\hat{IE}	0.551	0.368	0.296	0.227	0.209	0.215
Sample size: 1000, Range: [0.393, 7.438], $IE_{\mathcal{D}_{sample}} = 0.213$						
k	1	2	3	10	20	30
\hat{IE}	1.416	0.379	0.287	0.234	0.226	0.223

5 Conclusions and Future Work

We have introduced a novel non-parametric kNN approximation method for computing the IE from data samples. In accordance to our results, the $\hat{IE}_k^{(n)}(f)$ estimator is consistent.

In order to study the interaction between two random variables X and Y , the following measure of unilateral dependency was defined by Andonie *et al.* [1]:

$$o(Y, X) = IE(Y|X) - IE(Y)$$

This measure quantifies the unilateral dependence characterizing Y with respect to X and corresponds to the amount of information detained by X about Y . There is an obvious analogy between $o(Y, X)$ and the MI, since both measure the same phenomenon. However, the MI is a symmetric, not a unilateral measure.

Rather than approximating $o(Y, X)$ as we did in our previous studies, in our future work we will approximate directly the IE from the available dataset, using the $\hat{IE}_k^{(n)}(f)$ estimator. We also plan to apply our IE estimator to machine learning techniques.

Bibliography

- [1] Andonie, R.; Petrescu, F.; Interacting systems and informational energy, *Foundation of Control Engineering*, 11:53-59, 1986.
- [2] Andonie, R.; Cațaron, A.; An informational energy LVQ approach for feature ranking, *Proc. of the European Symposium on Artificial Neural Networks ESANN 2004, Bruges, Belgium, April 28-30, 2004*, D-side Publications, 471-476, 2004.
- [3] Andonie, R.; How to learn from small training sets, *Dalle Molle Institute for Artificial Intelligence (IDSIA)*, Manno-Lugano, Switzerland, September, invited talk, 2009.
- [4] Bonachela, J.A.; Hinrichsen, H.; Munoz, M.A.; Entropy estimates of small data sets, *J. Phys. A: Math. Theor.*, 41:202001, 2008.
- [5] Cațaron, A.; Andonie, R.; Energy generalized LVQ with relevance factors, *Proc. of the IEEE International Joint Conference on Neural Networks IJCNN 2004*, Budapest, Hungary, July 26-29, 2004, ISSN 1098-7576, 1421-1426, 2004.
- [6] Cațaron, A.; Andonie, R.; Informational energy kernel for LVQ, *Proc. of the 15th Int. Conf. on Artificial Neural Networks ICANN 2005, Warsaw, Poland, September 12-14, 2005*, W. Duch et al. (Eds.): Lecture Notes in Computer Science 3697, Springer-Verlag Berlin Heidelberg, 601-606, 2005.
- [7] Cațaron, A.; Andonie, R.; Energy supervised relevance neural gas for feature ranking, *Neural Processing Letters*, 1(32):59-73, 2010.
- [8] Cațaron, A.; Andonie, R.; How to infer the informational energy from small datasets, *Proc. of the Optimization of 13th International Conference on Electrical and Electronic Equipment (OPTIM2012)*, Brasov, Romania, May 24-26, 1065-1070, 2012.
- [9] Faivishevsky, L.; Goldberger, J.; ICA based on a smooth estimation of the differential entropy, *Proc. of the Neural Information Processing Systems, NIPS 2008*.

-
- [10] Gamez, J.E.; Modave, F.; Kosheleva, O.; Selecting the most representative sample is NP-hard: Need for expert (fuzzy) knowledge, *Proc. of the IEEE World Congress on Computational Intelligence WCCI 2008*, Hong Kong, China, June 1-6, 1069-1074, 2008.
- [11] Guiasu, S.; *Information theory with applications*, McGraw Hill, New York, 1977.
- [12] Hogg, R.V.; *Introduction to mathematical statistics, 6/E*, Pearson Education, ISBN 9788177589306, 2006.
- [13] Kraskov, A.; Stögbauer, H.; Grassberger, P.; Estimating mutual information, *Phys. Rev. E*, American Physical Society, 6(69):1-16, 2004.
- [14] Kozachenko, L. F.; Leonenko, N. N.; Sample estimate of the entropy of a random vector, *Probl. Peredachi Inf.*, 2(23):9-16, 1987.
- [15] Lohr, H.; *Sampling: Design and analysis*, Duxbury Press, 1999.
- [16] Miller, M.; Miller M.; *John E. Freund's mathematical statistics with applications*, Pearson/Prentice Hall, Upper Saddle River, New Jersey, 2004.
- [17] Onicescu, O.; Theorie de l'information. Energie informationelle, *C. R. Acad. Sci. Paris, Ser. A-B*, 263:841-842, 1966.
- [18] Paninski, L.; Estimation of entropy and mutual information, *Neural Comput.*, MIT Press, Cambridge, MA, USA, ISSN 0899-7667, 6(15):1191-1253, 2003.
- [19] Principe, J. C.; Xu, D.; Fisher, J. W. III.; Information-theoretic learning, *Unsupervised adaptive filtering*, ed. Simon Haykin, Wiley, New York, 2000.
- [20] Silverman, B.W.; *Density Estimation for statistics and data analysis (Chapman & Hall/CRC Monographs on statistics & Applied Probability)*, Chapman and Hall/CRC, 1986.
- [21] Singh, H.; Misra, N.; Hnizdo, V.; Fedorowicz, A.; Demchuk, E.; Nearest neighbor estimates of entropy, *American Journal of Mathematical and Management Sciences*, 23:301-321, 2003.
- [22] Walters-Williams, J.; Li, Y.; Estimation of mutual information: A survey, *Proc. of the 4th International Conference on Rough Sets and Knowledge Technology*, RSKT 2009, Gold Coast, Australia, July 14-16, 2009, Springer-Verlag, Berlin, Heidelberg, 389-396, 2009.
- [23] Wang, Q.; Kulkarni, S. R.; Verdu, S. (2006); A nearest-neighbor approach to estimating divergence between continuous random vectors, *Proc. of the IEEE International Symposium on Information Theory*, ISIT 2006, Seattle, WA, USA, July 9-14, 2006, 242-246, 2006.

Feature Clustering based MIM for a New Feature Extraction Method

S. El Ferchichi, S. Zidi, K. Laabidi, M. Ksouri, S. Maouche

Sabra El Ferchichi*

1. University of Tunis EL Manar
National Engineering School of Tunis
Tunisia, BP 37, LE BELVEDERE 1002, TUNIS
2. Lille1 University, Science and Technology
France, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex
*Corresponding author: sabra.elferchichi@enit.rnu.tn

Salah Zidi, Salah Maouche

Lille1 University, Science and Technology
France, Cité Scientifique, 59655 Villeneuve d'Ascq Cedex
salah.zidi@univ-lille1.fr, salah.maouche@univ-lille1.fr

Kaouther Laabidi, Moufida Ksouri

University of Tunis EL Manar, National Engineering School of Tunis, Tunisia, BP 37, LE BELVEDERE 1002, TUNIS kaouther.laabidi@enit.rnu.tn, moufida.ksouri@enit.rnu.tn

Abstract: In this paper, a new unsupervised Feature Extraction approach is presented, which is based on feature clustering algorithm. Applying a divisive clustering algorithm, the method search for a compression of the information contained in the original set of features. It investigates the use of Mutual Information Maximization (MIM) to find appropriate transformation of clusterde features. Experiments on UCI datasets show that the proposed method often outperforms conventional unsupervised methods PCA and ICA from the point of view of classification accuracy.

Keywords: feature extraction, Mutual Information Maximization (MIM), similarity measure, clustering.

1 Introduction

The capabilities of a classifier are ultimately limited by the quality of the features in each input vector. Using a large number of features can be wasteful of both computational and memory resources. Additionally, there are irrelevant and redundant features that complicate the learning process, and can lead to inaccurate prediction. Although those features may contain enough information about the output class, they can not predict the output label correctly because of the large dimension of the feature space and the reduced number of collected instances. It is important to note that for the classifier, it becomes more difficult to determine the inherent relation between the features and the class distribution [9]. This problem is commonly referred to as the curse of dimensionality [6].

A reduction of the feature space dimensionality is often necessary to alleviate this problem. To address this issue, two different approaches exist : Feature Selection which consists in selecting only the attributes which are relevant according to a pre-defined criterion [10]; And Feature Extraction which transforms the original set of feature and constructs a new one, more compact and more useful for the classification [20].

Feature Extraction methods like PCA [15], ICA [12] and Feature Selection methods, try either to find new statistically independent directions, or to eliminate totally the redundant features. An alternative approach is to gather the "similar" features into a much smaller number of feature-clusters, and use them to re-describe the data. Consequently, the potential information contained

in these features could be preserved while the size of the feature space is reduced and good performances are maintained. The crucial step in such a procedure is the characterization of the "similarity" between features. Recently, the use of clustering has been investigated for the extraction of features. The applicability of this approach was proven in the case of text classification problems [17], [1] and protein sequences analysis [4]. For each application domain, a specific functional similarity measures was determined.

In this work, we develop a new unsupervised Feature Extraction method. It is based on the use of clustering technique combined with Mutual Information Maximization (MIM) to perform feature clustering. Our main interest is to reveal the underlying structure of the feature space without any prior information about probabilities density functions or class-distribution of the data.

Usually, in high dimensional space there are many features that have similar tendencies along the dataset. They describe similar variations of monotonicity (increasing or decreasing trend). Those features give a related discriminative information for the learning process. Hence, an analysis of the variations of the monotonicity of each feature vector along the dataset can lead us to determine a form of redundancy in the data. By using trend analysis, each feature will be totally described by its signature, which is statistically distinguished from random behavior. Intuitively, once the groups of similar features have been settled, feature extraction can be realized through a linear or nonlinear transformation that will determine a representative feature for each feature-cluster. In the same time, the extraction has to preserve the main characteristics of each feature-cluster and to incorporate them into the new representative feature. Therefore, each feature has to be highly correlated with its corresponding group center. To satisfy this objective a reliable measure of dependency between each feature-cluster, its corresponding centroid and a search strategy are needed. Within this context, MI is a suitable dependency measure [19] for our problem: it quantifies the amount of information that the center carries about the feature-cluster. It can detect either a linear or a nonlinear relationship between two random variables [18], [16]. MI measure was exploited in feature extraction and selection method but in a supervised fashion [11], [20], [13] and [2].

In section 2, Feature Extraction based Clustering Method (FEMC) is briefly reviewed. In section 3, we focus on the formulation of the feature-cluster transformation, based on MI maximization (MIM). In section 4, the performances of the proposed Feature Extraction based Clustering Method (FEMC) is analyzed and discussed through multiple classification problems. In section 5, we offer some conclusions and suggestions for future work.

2 Feature Extraction Method based Clustering

The feature extraction method FEMC was recently proposed by [7] for pattern classification problems. It aims to obtain more generalization capabilities than existing methods. It performs feature extraction without presuming any knowledge about data structure or about instances' classes [7]. As we stated before, features that behave the same along the data may contain the same information. Grouping those features and transforming them guarantee there is no loss of information, better classification accuracy and reduced dimension. Hence, we focus on identifying "similar" features in their tendencies along the dataset. The analysis of features monotonicity reveals a form of redundancy between these features.

Clustering technique was used to identify complex relationships between features and to discover the inherent data structure [8], [21]. A k-means algorithm based on a new similarity measure was performed.

Analyzing the tendency of feature vectors was proposed to identify the similarity between them. This new measure was designed to overcome the limitations of Euclidean distance, usually used

in clustering algorithms.

In fact, a trend is a semi-quantitative information, describing the evolution of the qualitative state of a variable in a time interval, by using a set of symbols such as {increasing, decreasing, steady} [5]. It was used with success for process monitoring and diagnosis [5].

The procedure of feature extraction proposed in [7] start by computing the first order derivative of a feature vector is and fixing its sign (0, 1 or -1), at each point sample. After coding each trend, the difference of the tendency between each two vectors is computed. The distance is expressed as the squared root of the sum of the absolute difference between the occurrences of a specified value of a trend for two feature vectors. This was inspired from the Value Difference Metric (VDM) [14]. Thus, the location of a feature vector within the feature space is not defined directly by the values of its components, but by the conditional distributions of the extracted trend in each component. Furthermore, the similarity measure is not affected by the ordering of samples.

3 MIM for Feature Extraction based on Clustering

We consider $\{x_1, x_2, \dots, x_L\}$ the D -dimensional original dataset composed of D features v_j each. A clustering technique is performed on the feature space to construct $d < D$ feature-clusters. We have to define an appropriate transformation for each feature-cluster in order to obtain a representative features g_k , defined by the equation 1:

$$g_k = f(C_k) = \sum_{h=1}^{n_k} \sum_{i=1}^L w_i v_h, \tag{1}$$

Where, C_k is the cluster composed of n_k feature vectors v_h . w_i is the weight attributed to each component of the feature vector v_h . The transform w_i has to preserve any linear or nonlinear dependency between features in $v_h \in C_j$ and their centroid g_j .

MI is an appropriate measure of dependency, so the optimal transform W^* has to maximize the MI between $\{V_j, g_j\}$. V_j is the matrix containing the n_k feature vectors belonging to a cluster C_k . W is the vector containing the n_k weights w_i .

3.1 Mutual Information

Information theory provides the possibility to measure the information with MI [9], [20]. Let $p(x)$ and $p(y)$ be the probability density function (pdf) for random vector X and Y , and $p(x, y)$ the joint pdf. The MI between the discrete random vectors X and Y is defined as:

$$I(x, y) = \sum_{x \in \mathbf{X}} \sum_{y \in \mathbf{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{2}$$

Where \mathbf{X} and \mathbf{Y} are the corresponding alphabets of X and Y .

If the MI between the two random vectors is large then, the two vectors are closely related. If the MI becomes zero then, the two random vectors are independent. MI for continuous random variables are defined as follows:

$$I(X, Y) = - \int \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dx dy. \tag{3}$$

The determination of the pdfs ($p(x, y), p(x), p(y)$) and the performance of the integrations is very complicated. Consequently, the continuous input feature space is divided into several discrete

partitions. MI is then calculated using its expression for the discrete random variables. The inherent error that exists in the quantization process poses a problem. The Parzen window method is then used to estimate the pdfs of continuous random variables [19].

The method places a kernel function on top of each sample and evaluate the density as a sum of the kernels.

Given a data of n N -dimensional training vectors $\{x_1, \dots, x_n\}$, the pdf estimated by the Parzen window method is expressed by:

$$\hat{p}(x) = \frac{1}{n} \sum_{i=1}^n \Phi(x - x_i, \sigma I). \quad (4)$$

where $\Phi(\cdot)$ is the Gaussian window function given by

$$\Phi(x, \Sigma) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} x^T \Sigma^{-1} x\right) \quad (5)$$

where Σ is a covariance matrix of an N -dimensional random vector Z .

Quadratic Mutual Information: when the aim is not to compute an accurate value of the entropy of a particular distribution, but rather to find a distribution that maximizes or minimizes the entropy given some constraints, a large number of alternative entropy measures are produced [19].

One of these is the following continuous density:

$$D(f, g) = \int x(f(x) - g(x))^\alpha dx. \quad (6)$$

Since MI is expressed as the divergence between the joint density and the product of the marginal, we can insert this into the relation (6) and this way, the quadratic MI measure between two continuous variables X_1 and X_2 can be derived:

$$I(X_1, X_2) = \int \int (p(x_1, x_2) - p(x_1)p(x_2))^2 dx_1 dx_2. \quad (7)$$

3.2 Problem Formulation

As we stated before, our objective is to realize an appropriate transformation on each feature-cluster. Each clusters' center is usually computed as the bary-center derived by the equation (1); where $W_j = \mathbf{1} \frac{1}{n_j}$.

We look for a more appropriate transformation W^* , since the center is the representative feature of its cluster and it will be used as a new feature.

Since our objectif is to maximize the MI between each cluster C_j and its corresponding center g_j , we define the transformation f to apply on each feature vector $v_i \in C_j$, by $g_{ij} = f(w, v_i)$, which maximizes $I(g_j, V_j)$ (MI between $v_i \in C_j$ and g_j) as described in figure 1.

By using(7) we obtain:

$$I(g_j, V_j) = \int \int (p(g_j, v) - p(g_j)p(v))^2 dg_j dv. \quad (8)$$

We have to develop $p(g_j, v)$ to be able to compute (8).

Since the center $g_{ij} = f(w, v_i)$, belongs to the cluster C_j : $g_j \in C_j$, we get the final set of features:

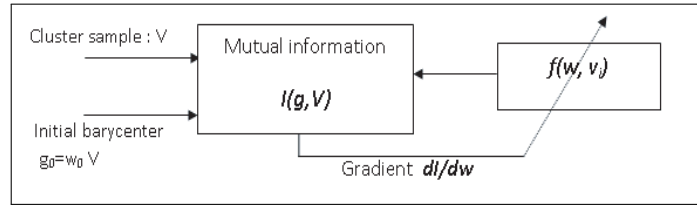


Figure 1: Feature Extraction procedure

$$S_{final} = V_j \cup \{g_j\}.$$

$p(g_j, v)$ will be expressed by:

$$p(g_j, v) = p(g_j)p(g_j/v) = p(g_j)(p(v) - p(g_j)). \quad (9)$$

We insert (9) in equation (8) and we get:

$$I(g_j, V_j) = \int \int p(g_j)^4 dg_j dv. \quad (10)$$

We have used the Parzen window estimator to determine $p(g_j)$. By using the equation (5) for the obtained set constituted of $n_j + 1$ features $S_f = V_j \cup \{g_j\}$, the density $p(g_j, v)$ can be expressed by:

$$p(g_j, v) = \frac{1}{n+1} \sum_{i=1}^{n+1} \Phi(g_j - v_j, \sigma I). \quad (11)$$

We know that $\int Z_c N_x(\mu_c, \Sigma_c) dx = Z_c$.

Henceforth, the MI becomes:

$$I(g, V) = \int_g \int_v \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} \sum_{l=1}^{n+1} \prod_{s=1}^{n+1} 4\Phi_g(\mu_s, \Sigma_s) dg dv \quad (12)$$

$$= \int_g \int_v \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} \sum_{l=1}^{n+1} z \Phi_g(\mu, \Sigma) dg dv \quad (13)$$

$$= \int_v \sum_{i=1}^{n+1} \sum_{j=1}^{n+1} \sum_{k=1}^{n+1} \sum_{l=1}^{n+1} z dv. \quad (14)$$

Where

$$z = \frac{|2\pi\Sigma_d|^{\frac{1}{2}}}{\prod_{s=1}^4 |2\pi\Sigma_d|^{\frac{1}{2}}} \prod_{a<b} \exp\left(-\frac{1}{2}(\mu_a - \mu_b)^T B_{ab}(\mu_a - \mu_b)\right),$$

$$\Sigma_d = \left(\sum_{s=1}^4 4\Sigma_s^{-1}\right)^{-1},$$

$$\Sigma_s = \sigma^2 I,$$

$$B_{ab} = \Sigma_s^{-1} \Sigma_d \Sigma_s^{-1}$$

In figure 1, we have to maximize the MI $I(g, v)$ to identify the optimal W^* and to determine cluster center g . We have used the gradient descent algorithm to find the solution which involves differentiating $I(g, v)$:

$$\frac{dI}{dw} = \frac{dI}{dg} \frac{dg}{dw} = \frac{dI}{dg} V. \quad (15)$$

Table 1: Classification accuracy on Sonar Dataset

No. of features	Classification accuracy (SVM)			
	PCA	ICA	FEMC (without MI)	FEMC (with MI)
1	58.6	67.2	60.60	65.94
3	54.7	69.7	71.68	75.05
6	63.0	70.2	77.39	77.89
9	70.2	68.7	82.38	82.32
12	75.1	71.7	85.02	81.40
60	82.7			

Table 2: Classification accuracy on Pima Dataset

No. of features	Classification accuracy (SVM)			
	PCA	ICA	FEMC (without MI)	FEMC(with MI)
1	66.3	73.2	68.08	67.80
2	75.1	76.7	69.72	69.85
3	75.5	76.8	75.72	75.72
5	75.5	77.2	76.80	82.32
60	78.0			

4 Experimental Results

In this section, we have conducted FEMC (with and without using MI) for different benchmark datasets from UCI machine learning repository [3]. We have compared the FEMC performances with conventional unsupervised Feature Extraction methods PCA and ICA for different extracted features. We have used SVM (given in Matlab toolbox) for binary classification task, the used kernel function is Gaussian kernel and the parameter σ is set after various tests. The K Nearest Neighbors (KNN) classifier is used for the multi-classification problem.

Sonar Dataset. We have used 13 fold cross-validation in experiments, as presented in [13]. For SVM parameters, we have set $\sigma = 1$. The Table 1 shows classification accuracy for different number of extracted features. The performances of FEMC are far better than PCA and ICA except for the case when the dimension is 1, and ICA outperforms the others. Since the concept of our approach is to form groups of similar features; extracting a very low number of features means gathering all features in a few numbers of clusters. This could be delicate for some datasets. We note also that in the case of dimension 9 and 12, FEMC can get nearly by the initial accuracy rate of 82% which is far better than ICA and PCA. By using MI, FEMC reaches much better accuracy, especially for the case of dimension 1, where it gets almost the same accuracy as ICA. Hence, MI increases the FEMC accuracy in the lower dimension like 1 and 3.

Pima Indian Diabetes Dataset. We have applied PCA, ICA and FEMC for comparison. A 10-fold cross strategy was used and $\sigma = 10$. Results are shown in Table 2. We can note that the classification accuracy of PCA and ICA becomes closer as the number of extracted features becomes larger. FEMC performs better than PCA and approaches the ICA accuracy for the dimensions: 3 and 5. But ICA still outperforms both PCA and FEMC for different numbers of features especially for the lower ones (dimension 1 and 2). By using MI, the accuracy of FEMC has increased especially for the case of the dimension 5, where it outperforms PCA, ICA and surpasses the initial accuracy (78%) by getting 82.32%.

Table 3: Classification accuracy on Breast Cancer Dataset

No. of features	Classification accuracy (SVM)			
	PCA	ICA	FEMC (without MI)	FEMC(with MI)
1	85.8	85.1	96.72	96.86
2	94.7	90.3	96.57	96.71
3	95.9	91.3	94.71	94.28
6	96.6	94.3	85.11	85.11
9	96.6			

Table 4: Classification accuracy on Ionosphere Dataset

No. of features	Classification accuracy (SVM)			
	PCA	ICA	FEMC(without MI)	FEMC(with MI)
1	64.07	61.28	72.09	76.09
3	85.21	81.80	75.20	83.50
6	84.79	86.05	85.21	85.49
9	84.83	86.52	87.52	87.80
12	86.31	88.04	89.20	89.20
34	91.73			

Breast Cancer Dataset. A 10-fold cross-validation was used and $\sigma = 0.01$. Results of comparison are shown in Table 3. With only one extracted feature, FEMC can get the maximum classification accuracy (96.86%). So, for a larger number of extracted features, PCA outperforms both ICA and FEMC and gets the maximum classification accuracy with 6 features. In this case, MI slightly ameliorates FEMC performances.

Ionosphere Dataset. We have used a 10 fold cross-validation and $\sigma = 0.01$. The results of the comparison are shown in Table 4. With only one extracted feature, FEMC outperforms ICA and PCA. For larger numbers of extracted features, FEMC gets either similar or better performance than PCA and ICA, and achieves the best classification accuracy with 12 features. In lower dimension, FEMC with MI reaches higher accuracy of 76.09% (better than PCA and ICA) especially in dimension 1.

Wine Dataset. We have used a 10-fold cross validation strategy and the K-nearest-neighbors (KNN) classifier to conduct classification task. The classification results for each Feature Extraction method are summarized in Table 5. We must underline that the FEMC performances are far better than ICA and PCA for low dimensions. Although for the dimension 2 ICA outperforms FEMC, for larger dimension FEMC achieves the best classification accuracy and approaches the initial one of 80.27%. By using MI, FEMC can reache better accuracy especially for the dimension 3 where it gets 81.49% (better than the initial one).

5 Conclusion

This paper deals with the important problem of extracting relevant features for pattern classification. Often, Feature Extraction techniques trust in some robust criterion to search for a lower dimensional representation. However, the true structure of the data is unknown, it is inherently ambiguous what constitutes a good low dimensional representation. This makes it difficult to define an appropriate criterion. We suggest a new Feature Extraction approach incorporating the idea of feature clustering. Similar features are recognized through analyzing their tendencies along the data set and a new similarity measure is then devised. The proposed approach FEMC

Table 5: Classification accuracy on Wine Dataset

No. of features	Classification accuracy (SVM)			
	PCA	ICA	FEMC(without MI)	FEMC(with MI)
1	67.93	67.42	71.02	71.88
2	71.94	73.84	72.36	76.25
3	72.48	75.22	78.19	81.49
5	75.74	91.01	79.92	80.45
13	80.27			

applies clustering technique, based on the new similarity measure, into feature space to determine its underlying groups of features. An MIM schema is used to find an optimal transformation of features in each obtained cluster to compute corresponding centers. The obtained set of centers represents the extracted features used to characterize the patterns.

The performances of FEMC method have been assessed through several datasets obtained from the UCI machine learning repository. The complexity of the relationships nature between features increases as the dimension gets lower. The MI can effectively identify this relationship due to its powerful background. In this work, we can clearly notice that by using MIM to construct a new set of features, FEMC have remarkably increased the classifier accuracy especially in lower dimension case. More, this method manage to offer a better accuracy classification than ICA and PCA in almost cases. The main problem that would occur is time complexity in estimating and optimizing MI between features for bigger datasets. As the optimization schema of MI is done separably for each cluster, a multi-agent approach can be useful to tackle this problem. The greedy algorithm used in this work can be also replaced by a stochastic one.

Bibliography

- [1] Baker, L.D. and McCallum, A.K.; Distributional clustering of words for text classification, *Proc. 21st Ann. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 1998.
- [2] Battiti R.; Using Mutual Information for Selecting Features in Supervised Neural Net Learning, *IEEE Trans. on Neural networks*, 5: 537-550, 1994.
- [3] Blake, C.L. and Merz C.J.; UCI repository of machine learning databases, <http://archive.ics.uci.edu/ml/>, Department of Information and Computer Science, University of California, Irvine, CA, 1998.
- [4] Bonet, I., Saeys, Y., Grau Abalo, R., García, M., Sanchez, R. and Van de Peer, Y. (2006); Feature extraction using clustering of protein, *Proc. 11th Iberoamerican Congress in Pattern Recognition CIARP*, eds. Springer, LNCS 4225, 614-623, 2006.
- [5] Charbonnier, S. and Gentil, S.; A trend-based alarm system to improve patient monitoring in intensive care units, *Control Engineering Practice*, 15:1039-1050, eds. Eds. Elsevier, Kidlington, ROYAUME-UNI, 2007.
- [6] Cherkassky, V. and Mulier, F.; *Learning from data: concepts, theory and methods*, chapter 5, eds. John Wiley & Sons, 1998.

-
- [7] EL Ferchichi, S., Zidi, S., Laabidi, K., Ksouri, M. and Maouche, S.; A new feature extraction method based on clustering for face recognition ,” *Proc. 12th Engineering Applications of Neural Networks*, eds. Springer, IFIP 363, 247-253, 2011.
- [8] Fern, X.Z. and Brodley, C.E.; *Cluster Ensembles for High Dimensional Clustering: an empirical study*, Technical report, CS06-30-02, 2004.
- [9] Fisher, J.W., Principe, J.C.; A methodology for information theoretic feature extraction, *Proc. 17th Int’l Joint Conf. on Neural Networks*, 1998.
- [10] Guyon, I. , Elisseeff, A.; An introduction to variable and feature selection, *Journal of Machine Learning Research*, 3: 1157-1182, 2003.
- [11] Hild II, K.E., Erdogmus, D., Torkkola, K., and Principe, J.C.; Feature extraction using information-theoretic learning, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28, 2006.
- [12] Kwak, N., and Choi, C.; Feature extraction based on ICA for binary classification problems, *IEEE Trans. on Knowledge and Data Engineering*, 15: 1387-1388, 2003.
- [13] Kwak, N., Feature selection and extraction based on mutual information for classification; Ph.D Thesis, Seoul National Univ., Seoul, Korea, 2003.
- [14] Payne, T.R. and Edwards, P.; Implicit feature selection with the value difference metric, *Proc. 13th European Conf. on Artificial Intelligence*, 1998.
- [15] Saul, L.K., Weinberger, K.Q., Sha, F., Ham, J. and Lee, D.D.; *Spectral Methods for Dimensionality Reduction, Semi supervised Learning*, eds. MIT Press Cambridge, MA, 2006.
- [16] Schaffernicht, E., Kaltenhaeuser, R.; On estimating mutual information for feature selection, *Proc. 17th Int’l Conf. on Artificial Neural Networks*, eds. Springer, LNCS 6352, 362-367, 2010.
- [17] Slonim, N. and Tishby, N.; The power of word clusters for text classification, *Proc. 23rd European Colloquim on Information Retrieval Research*, 2001.
- [18] Suzuki, T., Sugiyama, M., and Kanamori, T.; A Least-squares Approach to Mutual Information Estimation with Application in Variable Selection, *JMLR 17th 3rd Workshop on New Challenges for Feature Selection in Data mining and Knowledge Discovery (FSDM 2008)*, 2008.
- [19] Torkkola, K. and Campbell, W.M.; Mutual information in learning feature transformations, *Proc. 17th Int’l Conf. on Machine Learning*, 2000.
- [20] Torkkola, K., Feature extraction by non-parametric mutual information maximization, *Journal of Machine Learning Research*, 3: 1415-1438, 2003.
- [21] Von Luxburg, U., Bubeck, S., Jegelka, S. and Kaufmann, M.; Consistent minimization of clustering objective functions, *Neural Information Processing Systems NIPS*, 2007.

A Detailed Analysis of the GOOSE Message Structure in an IEC 61850 Standard-Based Substation Automation System

C. Kriger, S. Behardien, J. Retonda-Modiya

**Carl Kriger, Shaheen Behardien,
John-Charly Retonda-Modiya**

Centre for Substation Automation and Energy Management Systems
Cape Peninsula University of Technology
South Africa, P.O Box 1906, Bellville, Cape Town, 7535
krigerc@cput.ac.za, behardiens@cput.ac.za, retonda7@yahoo.fr

Abstract: In order to implement an IEC 61850 communication system, there needs to be a complete understanding of the methods, tools and technologies associated with the communication network, protocol and messaging underpinning the services. The IEC 61850 standard allows for communication between devices within a substation where a peer-to-peer model for Generic Substation Events (GSE) services is used for fast and reliable communication between Intelligent Electronic Devices (IEDs). One of the messages associated with the GSE services is the Generic Object Oriented Substation Event (GOOSE) message. A detailed analysis of the structure for the GOOSE message is required for fault diagnosis, or when developing hardware that is compliant with the IEC 61850 standard. This is one of the stated objectives of the Centre for Substation Automation and Energy Management Systems (CSAEMS) in the training of prospective specialists and engineers. A case study is presented where the structure of the GOOSE message as described in IEC 61850-8-1 is confirmed using firstly simulation, then experimentation with actual IEDs. In the first instance the message structure is confirmed by simulation of the GOOSE message and capturing it using network protocol analyzer software, after which analysis of the packet frame is performed. Data encoding of the GOOSE Protocol Data Unit (PDU) is analyzed with emphasis on the Abstract Syntax Notation (ASN. 1) Basic Encoding Rules (BER). The second part of the case study is conducted through experimentation with IEDs which are used to generate a GOOSE message and network protocol analyzer software is used to analyze the structure. Both the simulation and practical experimentation with actual devices confirm the GOOSE message structure as specified in part 8-1 of the IEC 61850 standard.

Keywords: IEC 61850, substation automation, Generic Object Oriented Substation Event (GOOSE), Intelligent Electronic Device (IED).

1 Introduction

The International communication standard for devices within a substation environment known as the International Electrotechnical Commission (IEC) IEC 61850 standard has contributed immensely to the way communication and information exchange are implemented within an electrical substation. This fairly new communication standard aims to ensure, amongst other things interoperability among devices from different vendors. For time-critical events such as the protection of electrical equipment, messages known as Generic Object-Oriented Substation Event (GOOSE) messages are exchanged between devices by means of a local Ethernet network. The paper presents a detailed examination and analysis of the captured GOOSE message structure generated firstly by means of simulation of the GOOSE message using software, then using actual protection devices.

The next section contains a brief examination of the GOOSE message followed by the simulation and practical case study setup. The GOOSE message frame is analyzed according to

the specification in IEC 61850-8-1 of the standard. The Application Protocol Data Unit and its relevance to the ASN. 1 Basic Encoding Rules is discussed and finally the conclusion and future prospects for this work are presented.

2 GOOSE Message Background

The devices in substations evolved from the older electromechanical relays into the current Intelligent Electronic Devices (IEDs) utilizing embedded microcontroller capabilities with communication between devices. The Ethernet technology was a natural part of this evolution. [1] With the advent of virtual local area networks (VLANs), an increase in data communication speeds, and flow control of switched systems, the Ethernet has become a reliable technology for this type of real-time application. [2]

The IEC 61850 standard allows for two groups of communication services between entities within the Substation Automation System (SAS), (IEC 61850-7-1) as shown in Fig. 1. [3] One group utilizes a client-server model, accommodating services such as Reporting and Remote Switching. The second group utilizes a peer-to-peer model for Generic Substation Event (GSE) services associated with time-critical activities such as fast and reliable communication between Intelligent Electronic Devices (IEDs) used for Protection purposes. In the IEC 61850-8-1 part of the standard, one of the messages associated with the GSE services are the Generic Object Oriented Substation Event (GOOSE) messages that allow for the broadcast of multicast messages across the Local Area Network (LAN). [4]

The Abstract Communications Service Interface (ACSI) shown in Fig. 1 and covered in greater detail in IEC 61850-7-2, defines common utility services for substation devices and shows the two groups of communication services for the client-server model and peer-to-peer model. The GSE model services provides a fast and reliable system-wide distribution of input and output data values; is based on a publisher/subscriber mechanism and, supports the distribution of the same generic substation event information to more than one physical device through the use of multicast/broadcast services (if so required and so engineered through the publish/subscribe mechanism).

Amongst others, two control classes and the structure of two messages are defined in IEC 61850: Generic Object Oriented Substation Event (GOOSE) supports the exchange of a wide range of possible common data organized by a DATA-SET; and Generic Substation State Event (GSSE) provides the capability to convey state change information (bit pairs). [5]

The scope of communication in IEC 61850 and the Logical interfaces are referenced in page 13 of the IEC61850-1 and give an indication that logical interface 8 is used for direct data exchange between bays (Peer-to-Peer), and GOOSE messages may also be passed between the bay and Station bus (level) as shown in Fig. 2. [6]

The Protocol Mapping Profile for the ACSI services in IEC 61850 is shown in Fig. 3. [4] The GOOSE message is associated with three layers of the Open Systems Interconnection (OSI) model, namely the Physical layer, Data-link layer and the Application layer. [4], [5] Page 114 of part IEC 61850-8-1 displays the structure of the GOOSE message and this is the starting point for the investigation and analysis of the GOOSE message structure. [4] To gain insight into the structure of GOOSE messages, a case study is presented in this paper in which a GOOSE message is simulated by means of the OMICRON IEDScout software, captured via network protocol analyzer software - Wireshark, and analyzed relative to the structure as defined in part 8-1

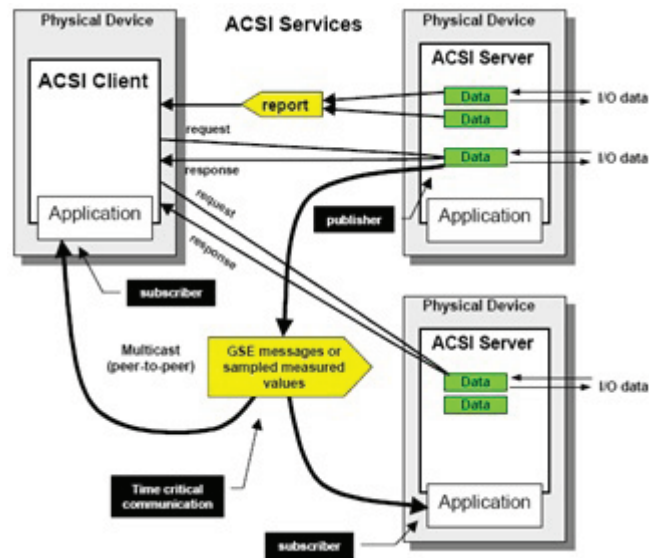


Figure 1: Abstract Communications Service Interface (ACSI) Services IEC 61850-7-1 pg 50 [3]

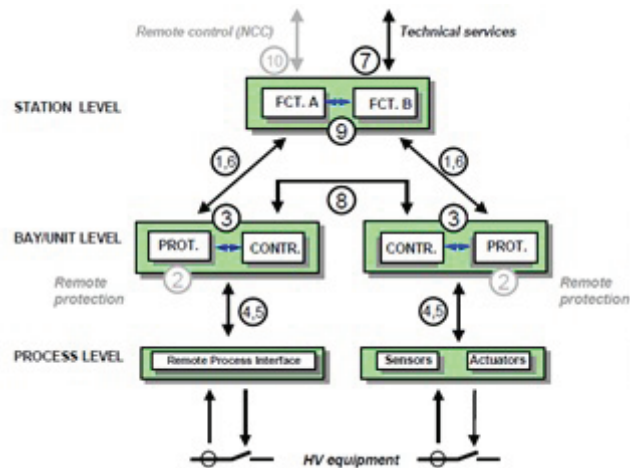


Figure 2: Interface Model of a Substation Automation System (pg 13 of [6])

pages 114-116 of the IEC61850 standard. [4] Further confirmation of the structure of a GOOSE message and the encoding of its data, is obtained through experimentation with actual IEDs from different vendors thus confirming interoperability as well.

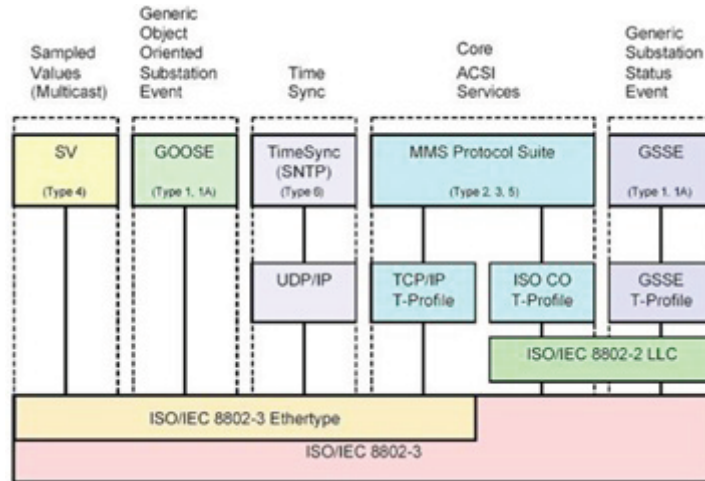


Figure 3: Communication profiles in IEC 61850 (pg 19 of [4])

3 Simulation of the GOOSE message and practical implementation

- Simulation study:

The OMICRON IEDScout software is an invaluable tool for educational purposes as it allows the simulation of the IEC61850 functionality such as GOOSE messaging (via OMICRON hardware platforms), and also has capabilities for monitoring real-time response on the network. The experimental setup for the GOOSE message simulation and validation is illustrated in Fig. 4 where IEDScout is used to generate a GOOSE message within the Personal Computer (PC) through the Network Interface Card (NIC) and Wireshark software running on a notebook is used to capture the GOOSE message packets. Both computers are connected to a network switch.

- Implementation study:

Fig. 5 shows the practical setup for confirmation of the GOOSE message structure by experimentation rather than simulation. IEDs from two different vendors are connected in a local area network via a network switch and GOOSE messages are exchanged between them. The test injection set injects 3-phase current into the Vendor no.1 IED. GOOSE messages are published from the Vendor no. 1 device and subscribed to by the Vendor no. 2 device. The measured 3-phase value currents are displayed on the Vendor no. 2 device front panel. To check for interoperability from both directions, the Vendor 2 device is configured as the publisher and the Vendor no.1 IED as the subscriber. A pushbutton on the Vendor 2 IED is pressed and a sequence of light emitting diodes (LEDs) on the Vendor no. 1 IED is lit. The data contained in the GOOSE messages being exchanged are binary and measurement data. The Wireshark software is used to confirm the structure of the captured message. [7]

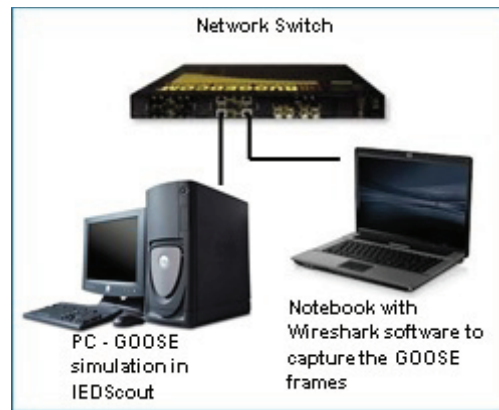


Figure 4: Case study - PC with IEDScout and PC with Wireshark Network Analyzer Software

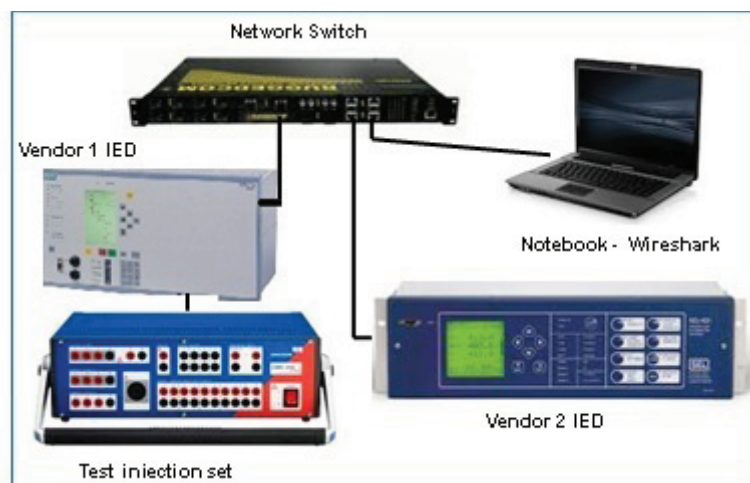


Figure 5: Practical experimental setup with IEDs from different vendors

4 Confirmation of the GOOSE Message Structure

Only the results of the GOOSE messages captured from the simulation case study in Fig. 4 are used in the next section for the message structure confirmation. A portion of the GOOSE message presented in the standard is shown in Fig. 6 on the LEFT hand side (red box). At the top of Fig. 6 (blue box) are the user-defined parameters for the GOOSE message generated by IEDScout software. To the bottom right of Fig. 6 (green box) is the Wireshark capture showing the hexadecimal values in the pane.

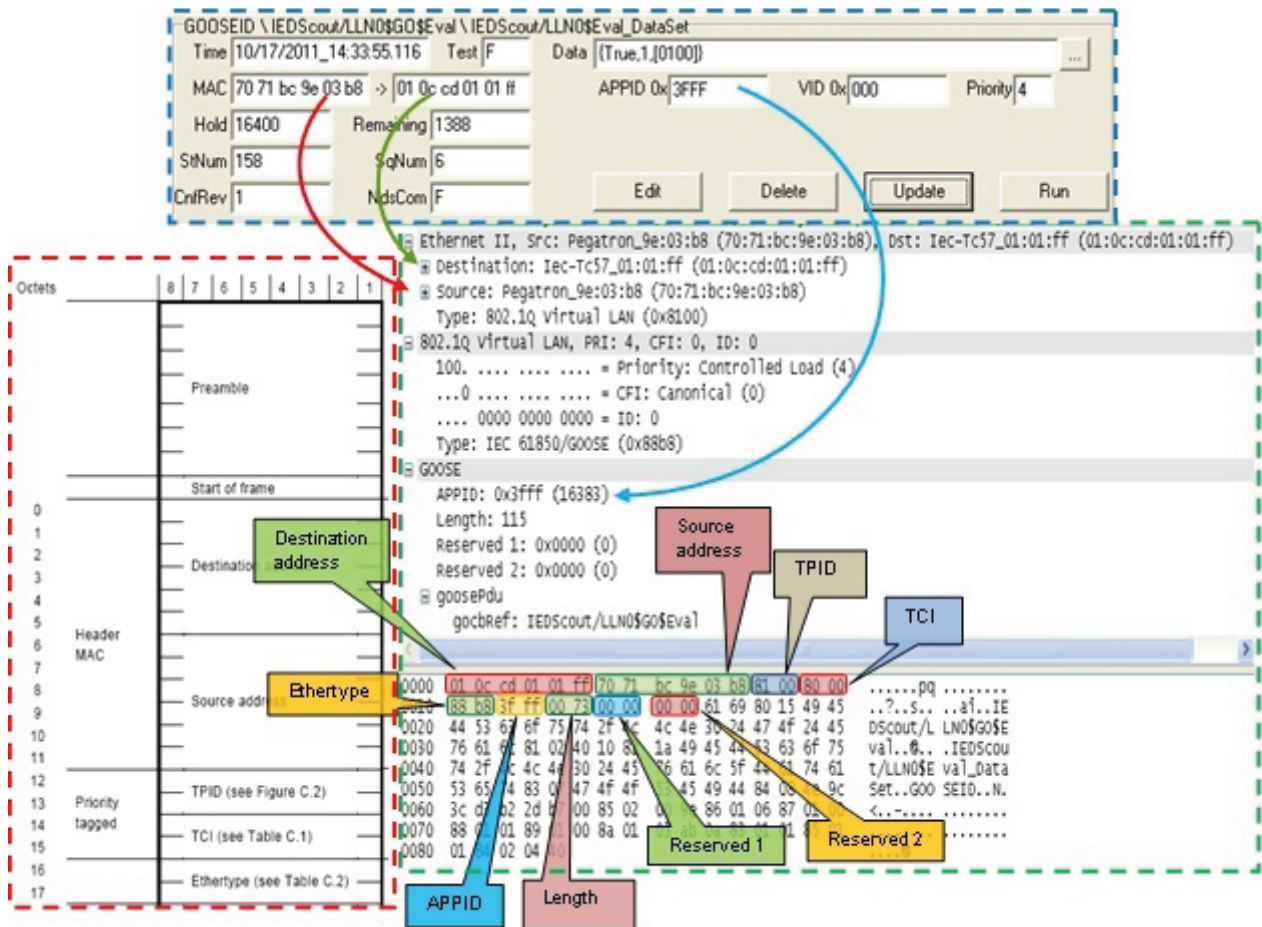


Figure 6: The FIXED section of the GOOSE message structure frame on the bottom left as defined in IEC 61850-8-1 pg 114 of [4]. The simulated GOOSE from IEDScout is at the top and the Wireshark window is at the bottom right

The GOOSE message structure consists of a portion that is fixed in terms of length, and fixed in terms of what content is specified to be there (entered via dialogue box) (Fig. 6). The next section of the message consists of a portion that is variable in terms of both length and content chosen to be communicated, e.g. the user defines which data elements and data attributes are to be transmitted (Fig. 10). Firstly the fixed portion of the message structure is examined and briefly discussed, starting with the Preamble and ending at the Etherbyte. The Preamble and Start of frame are performed at the hardware level. The Destination address is a multicast address consisting of 6 bytes. The Source address is also 6 bytes long. As per IEEE 802.1Q, priority tagging is used to separate time critical and high priority bus traffic for protection-relevant

applications. The 802.1Q Virtual Local Area Network (VLAN) is 4 bytes in length and consists of the TPID (Tag protocol identifier), TCI (Tag Control Information) and Ethertype. The TPID is the Ethertype assigned for 802.1Q Ethernet encoded frames and is given by 0x8100. The TCI consists of the CFI (Canonical Frame Indicator) and optional VID (VLAN Identifier). The TCI and Ethertype (0x88b8 for GOOSE) consists of 2 bytes each (IEC 61850-8-1 pg 115 . [4]). The application identifier (APPID) is 2 bytes in length and is used to select GOOSE messages from the frame and to distinguish the application association.

The discussion above indicates that all fixed components of the GOOSE message structure have been confirmed and accounted for. However, there were fields in the hexadecimal pane of the subsequent section in the message that warranted further investigation. As can be seen from Fig. 7, that when going from highlighted section Reserved 2 to goosePdu (GOOSE Data Protocol Unit), two bytes (with values 61 and 68) have been skipped. The result of this investigation pointed to another standard also referenced in IEC 61850-8-1, known as the Abstract Syntax Notation ONE, Basic Encoding Rules (ASN.1/BER) standard for Data networks and open system communications. [4] The next section considers aspects relating the ASN.1/BER standard to IEC 61850-8-1. The order in which elements occur in the GOOSE Protocol Data Unit (goosePdu) is always TAG, LENGTH followed by the DATA according to the ASN.1, as shown in Fig. 8. This will be explained in more detail later in this paper.

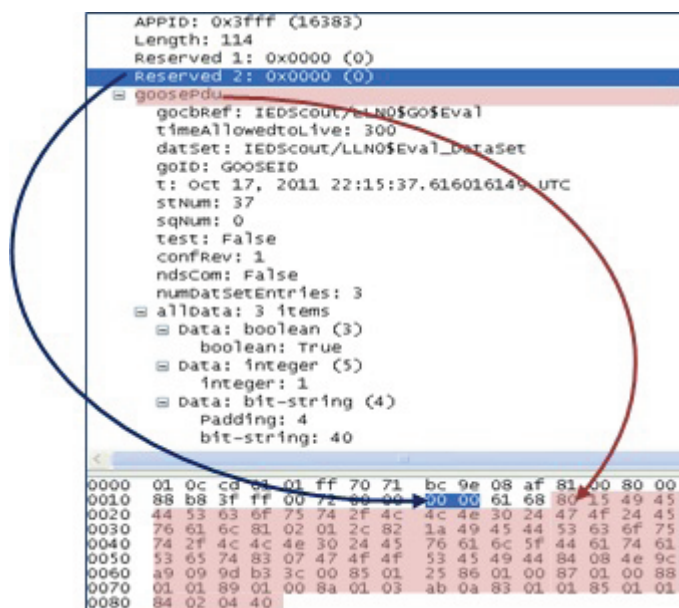


Figure 7: Indication of values skipped when transiting from "Reserved 2" to "goosePdu"

The variable portion of the message structure starts with the GOOSE Pdu Length (Fig. 8) up until the end of the message frame. The following 4 bytes are for the Reserved 1 and Reserved 2 fields. Following this is the Application Protocol Data Unit (APDU) in Fig. 9.

The APDU Length is 2 bytes long and indicates the size of the APDU with 8 octets added. The goosePdu in this case is 104 bytes in length and the GOOSE Control block reference (gocbRef) consists of 21 bytes. The timeAllowedtoLive is 2 bytes long and the data set of Logical node 0 is 26 bytes long. The GOOSE ID (goID) consists of 7 bytes, while the time (t) consists of 8 bytes. The status number (stNum), sequence number (sqNum), test bit, con-

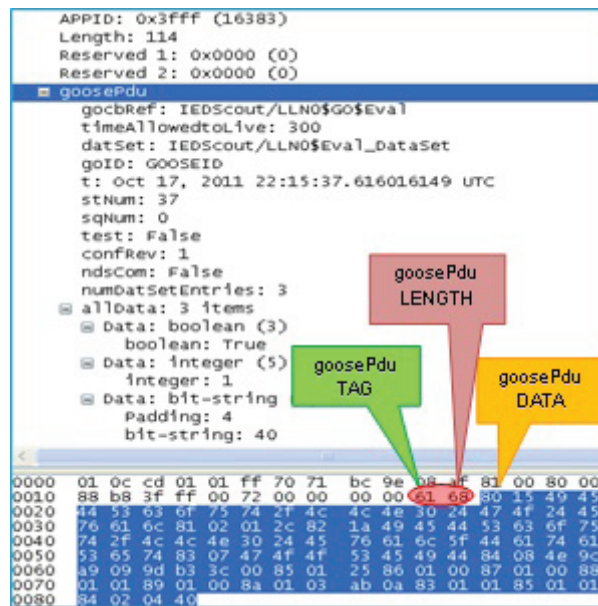


Figure 8: Illustrating the Length and Tag bytes and the contents of the GOOSE Protocol Data Unit

figuration revision (confRev), needs commissioning (ndsCom), and number of data set entries (numDatSetEntries) all are 1 byte in length. All these fields together with their functions are explained in detail in IEC 61850-8-1 page 114-116. [4] What is important to note is that before each of the preceding bytes, we first have the TAG, then the LENGTH followed by the actual DATA.

The last portion of the GOOSE message is the user-defined data content shown in Fig. 10. The user-defined data attributes in this particular case consists of three different items, namely a Boolean value, an integer and a data bit-string with padding. The Boolean and integer data items are 1 byte in length while the last data entry is 2 bytes long. The data section of the GOOSE message structure can be referenced in IEC 61850-7-2 page 116.] [8]

5 Application Protocol Data Unit (APDU) and ASN1 Basic Encoding Rules

The data within the GOOSE message are contained within the GOOSE Protocol Application Unit (PDU) and which is encoded in accordance with the Abstract Syntax Notation ONE (ASN. 1) standard for Data networks and open system communications (IEC 61850-8-1 pg 111). [4] The ASN.1 is an international standard used to define protocols of communication by means of encoding rules, IEC 61850-8-1, page111. The GOOSE protocol is defined using the ASN. 1/BER encoding rule (ASN. 1 Encoding Rule X.690-0207).

The encoding of GOOSE is not rigorously based on the original ASN. 1/BER but on an adaptation of ASN, 1/BER for Manufacturing Message Specification (MMS). It is important to note that this is not the original ASN.1 but a modified version as the original version does not cater for signed integers. [9], [10], [11]

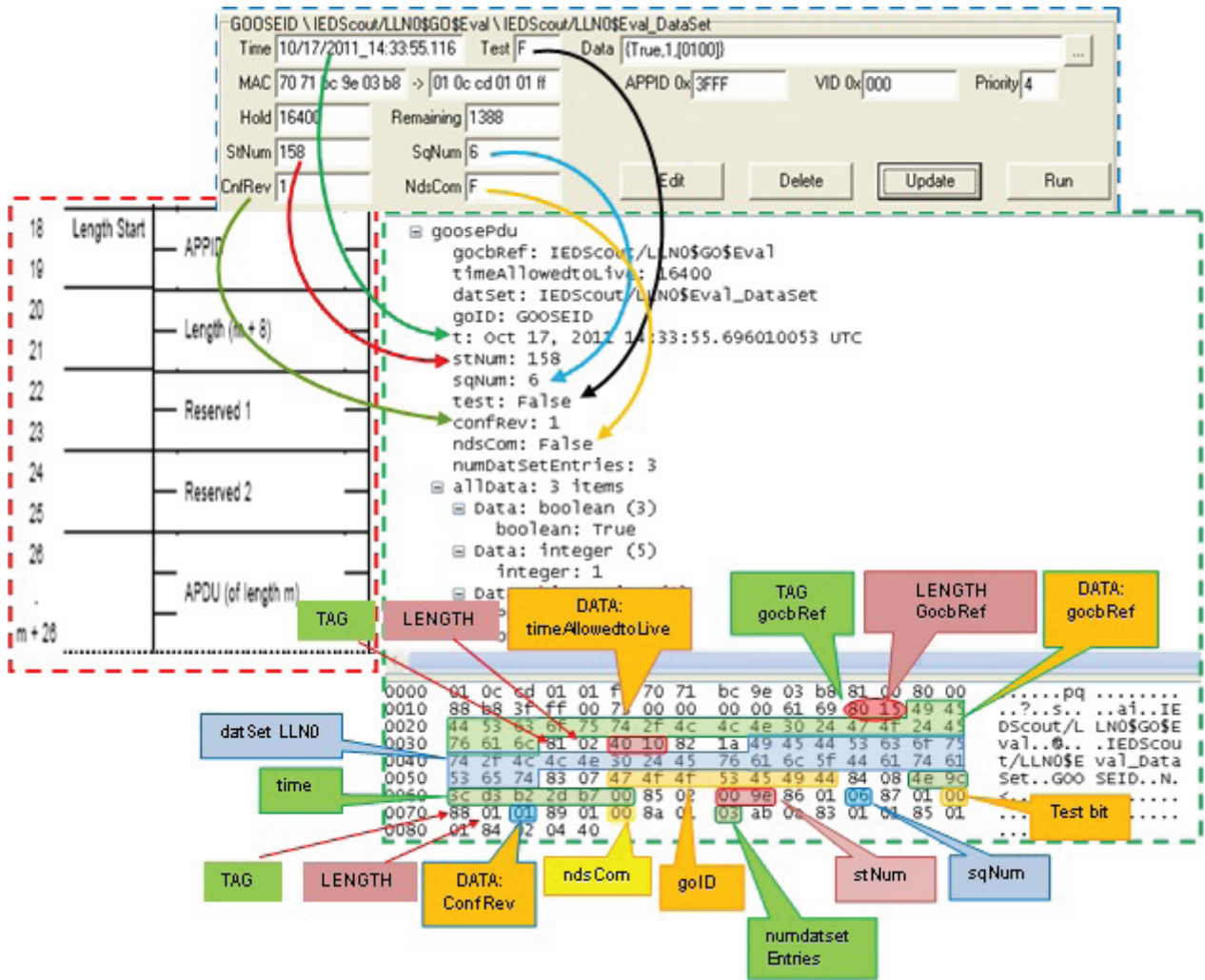


Figure 9: The VARIABLE section of the GOOSE message including the TAG, LENGTH, DATA order indicated by ASN. 1

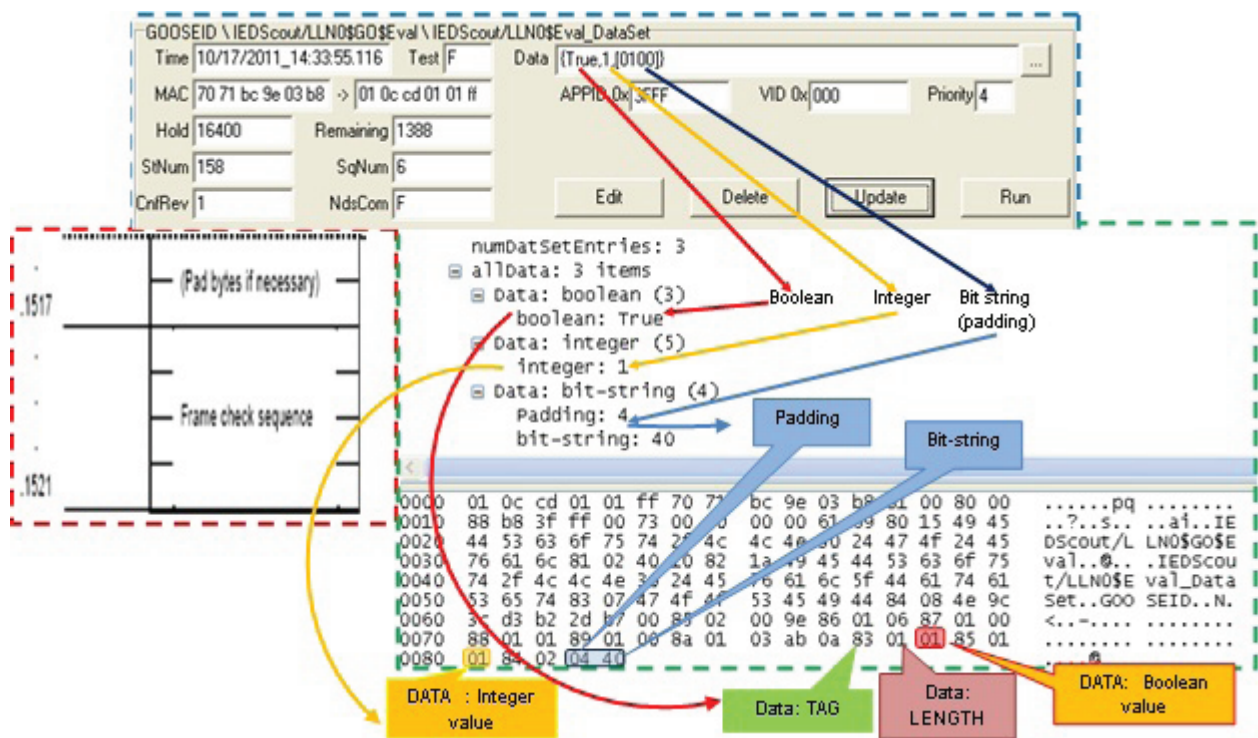


Figure 10: The user-defined dataset of the GOOSE message structure

The purpose of ASN.1 is to provide encoding and decoding specifications for protocol syntax that is to be sent over a network. The intent of this standard (ISO/IEC 8824 and 8825) is to have a neutral representation of fields as they are exchanged over a communication media. ASN.1 encoded values always have the format of TAG, LENGTH, followed by VALUE. [10] Fig. 11. The order in which each element appears within the GOOSE APDU (Application Protocol Data Unit) is also considered in ASN.1. [11]

Tag (1 byte)	Length	Content	End of content (optional)
--------------	--------	---------	---------------------------

Figure 11: An example of ASN.1 encoding showing the order of elements within the GOOSE APDU

The ASN.1 TAG describes the kind of information represented by the frame. The LENGTH is how many bytes (OCTETS) follow the DATA part of the frame. The LENGTH has a simple and extended form. If the length is less than 128 bytes, a single octet is used with its MSB set to 0. If the length is greater than 128, the MSB is set to 1, and the remaining 7 bits expressing the number of bytes that will contain the following parameter length. [11] The VALUE is the actual data content being specified by the frame; the value in the field must be interpreted according to the type of field. For example, if the field type is string, the octets must be decoded as characters, but if specified as an integer, the numeric value will be calculated based on the binary content.

The goosePdu always starts with a Tag (Identifier octet) containing the Class/Type of tag; the Primitive or Constructed Flag and the Tag number.

Description	Value octet (Hexadecimal)	Value octet (Binary)	Meaning (Interpretation)
Tag	61	01100001	Class APPLICATION, Composition of data types Tag number 1. Identifier octet. Fieldname "goosePdu"
Length	68	01101000	Length of "goosePdu"

Figure 12: Start of goosePdu tag header

With respect to Fig. 7, 8 and the table in Figure 12, it can clearly be seen that the bytes which were skipped refer to a class application type in this case goosePdu, and the associated length of the goosePdu. Similar common header tags appear throughout the PDU section of the frame. Some of these are presented in the tables displayed in Figures 13 and 14 including the tag headers, lengths and values for each byte within the goosePdu for the values in the frame represented in Figures 6 through 10.

The results of the simulation, practical implementation and the analysis thereof, confirm that the structure and data content for the GOOSE message is the same for both the experimentation and the simulation. The conclusion reached is that the structure of the GOOSE frame with respect to the IEC 61850-8-1 standard specified in page 114 and the content (user-defined data) of the GOOSE message with respect to IEC 61850-7-2 page 116 is confirmed.

6 Conclusions and Future Works

The IEC 61850 standard consists of many different parts and some are extremely difficult to understand and interpret even among domain experts [12]. Students also have to grapple with the challenges presented by the IEC 61850 standard and the approach adopted in this work was to develop a detailed understanding of the message structure communicated between the various devices within the substation network for instances requiring fault diagnosis to determine the cause of maloperation or interoperability issues [13].

The paper has discussed a detailed investigation and confirmation of the GOOSE message structure and data content through a process of simulation and experimentation. This was confirmed by referencing to the relevant parts of the IEC 61850 standard and also the modified ASN.1/BER standard. The process of confirmation of message structure and content has enabled a foundation to be established from which further educational activities such as embedded systems development and diagnostic activities might be pursued.

Acknowledgements

The research work described in the paper is partly funded by the National Research Foundation grant ICD2006061900021 and by the Eskom TESP program under the grants Investigation into Standard Based Substation Automation and Energy Management Systems, Methods for Power System State Estimation and GOOSE and Sampled Value Message structure investigation for IEC 61850 Standard-based implementation.

Description	Value octet (Hexadecimal)	Value octet (Binary)	Meaning (Interpretation)
Tag	80	10000000	Class CONTEXT-SPECIFIC, Primitive, tag number [0], IMPLICIT VISIBLE-STRING, Identifier octet. Fieldname "goose.gocbRef"
Length	15	00010101	Length of "goose.gocbRef"
Data	49 45 44 53 63 6F 75 74 2F 4C 4C 4E 30 24 47 4F 24 45 76 61 6C		Content of "goose.gocbRef" {IEDScout/LLN0\$GOSEval}
Tag	81	10000001	Class CONTEXT-SPECIFIC, Primitive, tag number [1], IMPLICIT INTEGER, Identifier octet. Fieldname "goose.timeAllowedtoLive"
Length	02	00000010	Length of "goose.timeAllowedtoLive"
Data	40 10	01000000 00010000	Content of "goose.timeAllowedtoLive" { 300 }
Tag	82	10000010	Class CONTEXT-SPECIFIC, Primitive, tag number [2], IMPLICIT VISIBLE-STRING, Identifier octet. Fieldname "goose.datSet"
Length	1A	00011010	Length of "goose.datSet"
Data	49 45 44 53 63 6F 75 74 2F 4C 4C 4E 30 24 45 76 61 6C 5F 44 61 74 61 53 65 74		Content of "goose.datSet" { IEDScout/LLN0\$Eval_DataSet }
Tag	83	10000011	Class CONTEXT-SPECIFIC, Primitive, tag number [3], IMPLICIT VISIBLE-STRING OPTIONAL, Identifier octet. Fieldname "goose.goID"
Length	07	00000111	Length of "goose.goID"
Data	47 4F 4E 53 45 49 44		Content of "goose.goID", { GOOSEID }
Tag	84	10000100	Class CONTEXT-SPECIFIC, Primitive, tag number [4], IMPLICIT UTCTime, Identifier octet. Fieldname "goose.t"
Length	08	00001000	Length of "goose.t"
Data	4E 9C 3C D3 B2 2D B7 00		Content of "goose.t", { 4c:39:04:ff:d0:e6:26:00 }
Tag	85	10000101	Class CONTEXT-SPECIFIC, Primitive, tag number [5], IMPLICIT INTEGER, Identifier octet. Fieldname "goose.stNum"
Length	02	00000001	Length of "goose.stNum"
Data	00 9E	00000000 10011110	Content of "goose.stNum", { 1 }
Tag	86	10000110	Class CONTEXT-SPECIFIC, Primitive, tag number [6], IMPLICIT INTEGER, Identifier octet. Fieldname "goose.sqNum"
Length	01	00000001	Length of "goose.sqNum"

Figure 13: Examination of each tag within the goosePdu for the given case study

Data	06	00000000	Content of "goose.sqlNum", { 0 }
Tag	87	10000111	Class CONTEXT-SPECIFIC, Primitive, tag number [7], IMPLICIT BOOLEAN DEFAULT FALSE, Identifier octet. Fieldname "goose.test"
Length	01	00000001	Length of "goose.test"
Data	00	00000000	Content of "goose.test", { 0 }
Tag	88	10001000	Class CONTEXT-SPECIFIC, Primitive, tag number [8], IMPLICIT INTEGER, Identifier octet. Fieldname "goose.confRev"
Length	01	00000001	Length of "goose.confRev"
Data	01	00000001	Content of "goose.confRev", { 1 }
Tag	89	10001001	Class CONTEXT-SPECIFIC, Primitive, tag number [9], IMPLICIT BOOLEAN DEFAULT FALSE, Identifier octet. Fieldname "goose.ndsCom"
Length	01	00000001	Length of "goose.ndsCom"
Data	00	00000000	Content of "goose.ndsCom", { 0 }
Tag	8A	10001010	Class CONTEXT-SPECIFIC, Primitive, tag number [10], IMPLICIT INTEGER, Identifier octet. Fieldname "goose.numDatSetEntries"
Length	01	00000001	Length of "goose.numDatSetEntries"
Data	03	00000011	Content of "goose.numDatSetEntries", { 3 }
Tag	AB	10001011	Class CONTEXT-SPECIFIC, Constructed, tag number [11], IMPLICIT SEQUENCE OF Data, Identifier octet. Fieldname "goose.allData"
Length	0A	00001010	Length of "goose.allData"
Tag	83	10000011	Class CONTEXT-SPECIFIC, Primitive, Boolean (3) Identifier octet. Fieldname "Data: Boolean"
Length	01	00000001	Length of "Data: boolean"
Data	01	00000000	Content of "Data: boolean ", { 0 }
Tag	85	10000101	Class CONTEXT-SPECIFIC, Primitive, Integer (5) Identifier octet. Fieldname "Data: integer"
Length	01	00000001	Length of "Data: integer "
Data	01	00000001	Content of "Data: integer ", {0}
Tag	84	10000100	Class CONTEXT-SPECIFIC, Primitive, bit-string (4) Identifier octet. Fieldname "Data: bit-string" "BER.bitstring.padding" "goose.bit_string"
Length	02	00000010	Length of "Data: bit-string"
Data	04 40	00000100 01000000	Content of "Data:Padding", {4} Content of "Data: bit-string", { 80 }

Figure 14: Examination of each tag within the goosePdu for the given case study (continued)

Bibliography

- [1] Dolezilek D.; IEC 61850: What you need to know about functionality and practical implementation, Power Systems Conference: *Advanced Metering, Protection, Control, Communication, and Distributed Resources*, PS 06 117, 2006.
- [2] De Oliveira J.C., Varella W.A., Marques A.E., Forster G.; Real Time Application using multicast Ethernet in Power Substation Automation according to IEC61850, *PAC World Journal*, 2007.
- [3] IEC 61850-7-1.; Communication networks and systems in substations - Basic communication structure for substation and feeder equipment Principles and models, *International Electrotechnical Commission (IEC)*, 2003.
- [4] IEC 61850-8-1.; Communication networks and systems in substations - Specific Communication Service Mapping (SCSM) Mappings to MMS (ISO 9506-1 and ISO 9506-2) and to ISO/IEC 8802-3, *International Electrotechnical Commission (IEC)*, 2003.
- [5] Apostolov A.; Fundamentals of IEC 61850, Seminar presented at the *Cape Peninsula University of Technology*, Cape Town, South Africa, 2009.
- [6] IEC 61850-1.; Communication networks and systems in substations Introduction and overview, *International Electrotechnical Commission (IEC)*, 2003.
- [7] Makhetha M., Kriger C.; Data acquisition and data distribution in an IEC 61850 standards-based substation environment. Unpublished thesis, Cape Peninsula University of Technology, Electrical Engineering Department, 2011.
- [8] IEC 61850-7-2.; Communication networks and systems in substations - Basic communication structure for substation and feeder equipment Abstract Communication Service Interface (ACSI), *International Electrotechnical Commission (IEC)*, 2003.
- [9] Retonda J., Behardien S.; Simulation of an IEC 61850 Based GOOSE Message, and Analysis of its Structure, *OMICRON Users Conference*, 2010.
- [10] Falk, H., Burns, M.; MMS and ASN.1 Encodings Simple Examples and Explanations on How to Crack an MMS PDU, *Systems Integration Specialists Company, Inc.(SISCO)*, USA, 1996.
- [11] Cesi Ricerca, Valutazione delle tempistiche associate ai messaggi di tipo GOOSE nellambito del protocollo IEC-61850, *TTD Tecnologie T and D*, 2006.
- [12] Liang, Y., Campbell, R.H.; Understanding and Simulating the IEC 61850 Standard. *Department of Computer Science University of Illonois. Urbana, USA*, 2008.
- [13] Tzoneva,R. Apostolov, A. Behardien, S. Kriger, C. Boesak, D. Gumede, C. ; IEC 61850 Standard Based Substation Automation Challenges To Universities. *PAC World Congress, Dublin Ireland*, 2010.

Fast and Accurate Home Photo Categorization for Handheld Devices using MPEG-7 Descriptors

B. Oh, J. Yu, J. Yang, J. Nang, S. Park

**Byonghwa Oh, Jungsoo Yu, Jihoon Yang,
Jongho Nang, Sungyong Park**

Department of Computer Science, Sogang University
1 Sinsu-dong, Mapo-gu, Seoul 121-742 Korea
mrfive@sogang.ac.kr, yjs@mlneptune.sogang.ac.kr,
yangjh@sogang.ac.kr, jhnang@sogang.ac.kr, parksy@sogang.ac.kr

Abstract:

Home photo categorization has become an issue for practical use of photos taken with various devices. But it is a difficult task because of the semantic gap between physical images and human perception. Moreover, the object-based learning for overcoming this gap is hard to apply to handheld devices due to its computational overhead. We present an efficient image feature extraction method based on MPEG-7 descriptors and a learning structure constructed with multiple layers of Support Vector Machines for fast and accurate categorization of home photos. Experiments on diverse home photos demonstrate outstanding performance of our approach in terms of the categorization accuracy and the computational overhead.

Keywords: machine learning, feature extraction, image classification, mobile computing, content based retrieval.

1 Introduction

Nowadays, there have been great advances in technology related to computers and cameras. People can easily take pictures anytime and anywhere using handheld devices such as cell/smart phones, digital cameras, game consoles, etc. As a result, the management of such *home photos* (taken by amateurs, rather than professionals) has become very important for their practical processing, storage, and use. Unfortunately, as mentioned in [1], home photos vary significantly unlike professional or domain-specific images, and the subjects in them are often misinterpreted. Therefore, browsing, searching, and categorizing such photos are nontrivial tasks.

We consider automatic categorization of home photos in this paper. Manual categorization is not appropriate since the time required for it can be even longer than that for the creation of a photograph. Moreover, people have different criteria for categorizing images, which produces non-uniform, unreliable results. So, it is of interest to develop an accurate, automatic categorization method for home photos.

Many researchers have proposed image categorization methods involving feature extraction and learning structures. Some of the studies include object-based learning and their spatial properties, focusing on the relationships among objects from a regional point of view [2, 3]. However, such method is not appropriate for handheld devices since object segmentation is very time-consuming. Thus, even faster and simpler extraction methods need to be devised instead of applying the region or object-based approaches.

The aim of this paper is to present a fast feature extraction method and an efficient learning structure suitable for accurate categorization of home photos, especially for handheld devices. For the former, we use a set of simple feature extractors in MPEG-7 descriptors [4] and a rapid face detector. For the latter, we present a hierarchical learning structure with Support Vector Machines (SVMs) [5] with the consideration of the meaning of concepts. Our approach is tested with a variety of real-world home photos and deployed on actual handheld devices, which

demonstrated good categorization capability. The rest of the paper is organized as follows: Section 2 defines the image features and introduces our feature extraction methods prior to learning. Then the overall structure of the learning method is described in Section 3. Section 4 explains the data, experimental setup, and the results. Section 5 concludes with a summary and discussion of some directions for future research.

2 Image Features and Their Extraction

In order to classify images by categories, we need to train classifiers taking feature inputs of the image and producing category outputs. The features can take various forms. Usually numeric values are used in training because many common classifiers act on numeric inputs for learning and making predictions [6]. We thus decided to use some of the MPEG-7 visual description methods, and an efficient face detector which detects the regions of frontal upright face objects in an image. The extracted values of all these features are numeric.

2.1 MPEG-7 Visual Descriptors

MPEG-7, an ISO/IEC standard developed by MPEG (Moving Picture Experts Group), provides a rich set of standardized tools to describe multimedia content [4]. It is able to efficiently search and retrieve relevant information that people want to use. There are several parts of standards in MPEG-7, and one of them, MPEG-7 visual, covers the following visual descriptors: Color, Texture, Shape, Motion, Localization, and Face recognition.

Eidenberger asserted that an efficient (general-purpose) descriptor should provide a surjective mapping from media object to points in feature space [7]. He supposed an ideal descriptor should be highly discriminating for any type of media content. He concluded that the best descriptors for combinations are Dominant Color Descriptor (DCD), Color Layout Descriptor (CLD), Edge Histogram Descriptor (EHD), and Texture Browsing Descriptor (TBD).

However, the extraction of texture descriptors such as TBD actually entails a higher time complexity than extracting color descriptors (DCD and CLD). This made us select the three best descriptors, DCD, CLD and EHD. Also, we added Color Structure Descriptor (CSD) to make up the exclusion of TBD and enhance the expressive power. From our previous research, we developed an extraction method for CSD which is much faster than other methods [8]. In the end, we used four MPEG-7 visual descriptors, CSD, DCD, CLD, and EHD. The first three descriptors are color descriptors and the last is a texture descriptor. Color descriptors have the ability to characterize the perceptual color similarity and generally have low complexities of extraction and matching. EHD characterizes the structures in generic images in forms of edge contents and layouts.

Prior to this work, we developed optimized versions of software engines that extract CSD, CLD and DCD [8]. Table 1 shows the computation time of some of the visual descriptors measured on handheld devices. Our engines were much faster than the XM reference software [9] which provides non-optimized extraction methods of visual descriptors. The speed gap will be even greater for images with lower resolutions (e.g., 320×240) instead of 640×480 . Additional descriptors such as Homogeneous Texture Descriptor (HTD) and Region Shape Descriptor (RSD) are very time-consuming and therefore they are unsuitable for use in handheld devices.

2.2 Face Descriptor

As the existence of particular objects in the image can aid the categorization process, we adopt the efficient face detector proposed in [10] which is trained with about 100,000 manually

Table 1: MPEG-7 Visual Descriptor Profiling Performances for 640×480 Images (in Milliseconds).

Descriptor	HP iPAQ rx5965 PDA (ARM9 400Mhz)		Samsung Omnia 2 Phone (ARM11 800Mhz)	
	XM Reference	Optimized	XM Reference	Optimized
CSD	4,500	900	2,800	850
CLD	150	50	50	30
DCD	23,000	170	13,000	110
EHD	600	-	550	-
HTD	9,600	-	8,400	-
RSD	61,000	-	50,550	-

cropped upright frontal face images. Then we detect at most 10 faces in an image each of which is represented by its area to define the Face Descriptor (FD).

3 The Process of Categorization

In order to develop the home photo categorization system, it is needed to train a classifier that classifies images under predefined categories. We build two-layered independent classifiers in order to enhance the classification performance in two steps. Figure 1 depicts the overall structure of the proposed categorization system.

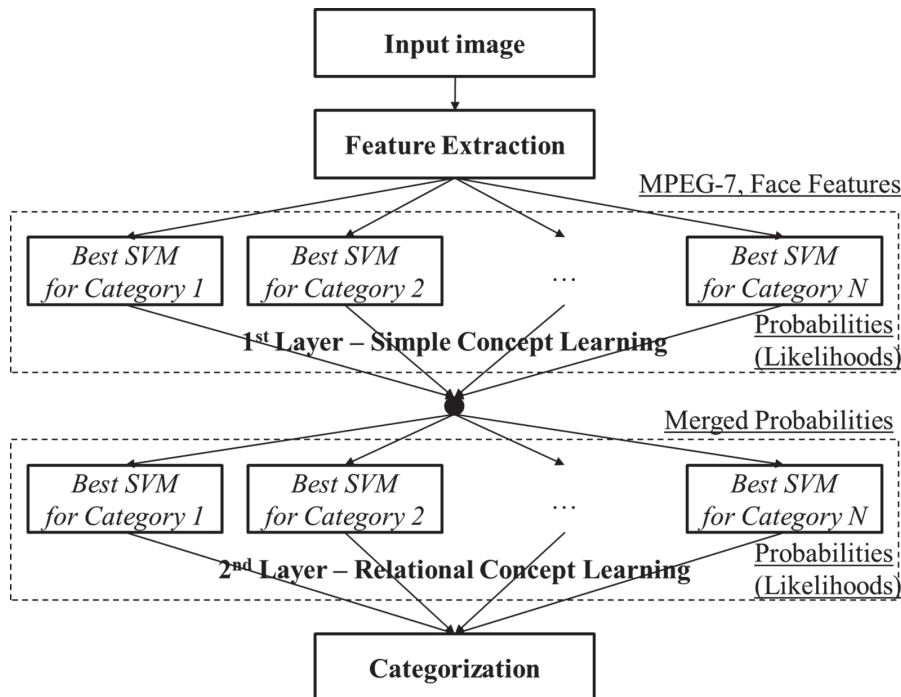


Figure 1: Overall Structure of Home Photo Categorization System.

First, the feature vectors are constructed using the MPEG-7 feature extractors and the face detector as described in Section 2. Then, the classifiers in the 1st layer are trained with the feature vectors to produce the probabilities (or likelihoods) for an image to belong to each category.

Next, the classifiers in the 2nd layer take the probabilities produced in the previous layer as inputs, and compute a new set of probabilities for each category, by considering the constraints immanent in the data. For instance, there may be constraints such as: “*If it is nighttime, then it may not be a landscape*” or “*Most of the photos taken of waterside regions are images of nature, not cities*”, and so on. Then the probability for a landscape will be lowered if the probabilities for nighttime and a landscape are both high, by passing through the 2nd layer. Since these constraints are unknown, the classifiers of the 2nd layer should be trained to reflect such knowledge. In other words, the 2nd layer is in charge of incorporating relational meanings among the categories.

Lastly, the system decides the category the image belongs to. For this, our system simply chooses the category with the greatest probability over all N values for N categories.

In order to follow this process, the best classifiers (in terms of certain performance criteria such as classification accuracy or F1 measure) need to be constructed in both layers.

3.1 Building the First Layer Classifiers

We adopt Support Vector Machines (SVMs) [5] as the base classifiers in our system, with four commonly used kernels (linear, polynomial, RBF, sigmoid). We also applied feature subset selection to obtain the best performance with minimum computational overhead and data acquisition cost. So with the five features of CSD, DCD, CLD, EHD and FD extracted in Section 2, we can construct 31 feature subsets (of different feature combinations).

Now we can find the best classifier for each category by changing kernels and feature subsets with 4×31 experiments for a given training data. This process is repeated to determine the best classifiers for all categories. Figure 2 illustrates the process.

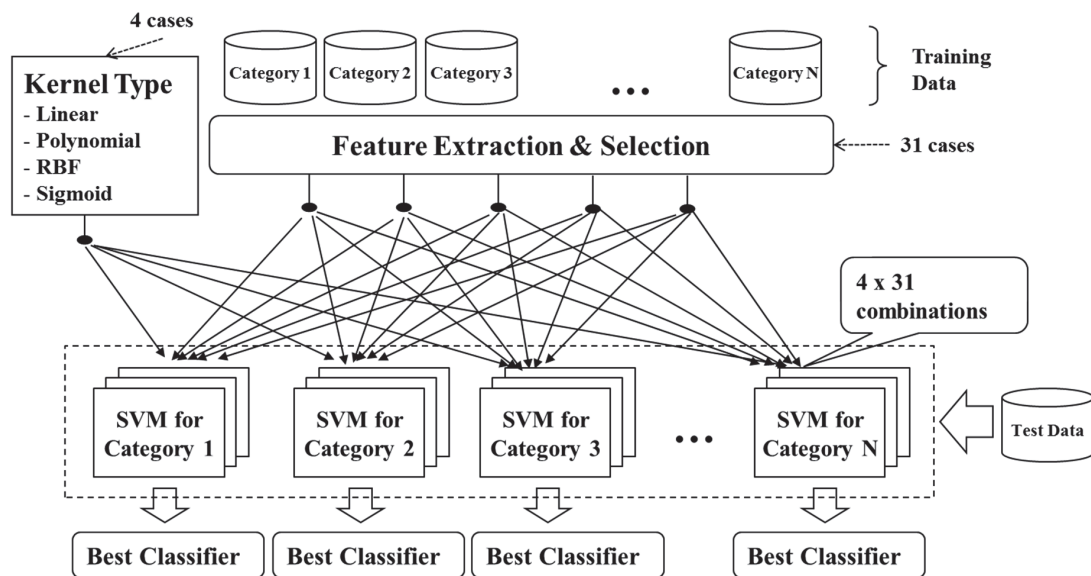


Figure 2: Process of Finding the Best Classifiers in the First Layer.

3.2 Building the Second Layer Classifiers

As a result of the first layer learning, each SVM produces the probability that an image belongs to the category it represents. But these outputs are not guaranteed to be correct since each probability is induced by separate SVMs without considering the correlation (or constraints)

Table 2: Discovered Rules on the Concept “When”.

Rule number (Importance)	Rule
1 (Strongest)	IF Night < 0.6625 AND Sunrise/Sunset < 0.4004 THEN Day
2	ELSE IF Landscape > 0.0024 AND Stadium < 0.04265 AND Evening \leq 0.5206 AND Night \leq 0.9865 AND Snow < 0.2154 AND Sunrise/Sunset \leq 0.9625 AND Waterside < 0.5083 THEN Day
3	ELSE IF Stadium < 0.2822 AND Day < 0.9575 AND Evening > 0.2788 AND Snow < 0.2154 AND Sunrise/Sunset \geq 0.3696 AND Waterside < 0.5083 THEN Evening
4 (Weakest)	ELSE Night

among the categories. For example, if there are more photos that capture nature in daylight than night, then there could be stronger correlation between daylight and nature than night and nature.

In order to check the utility of learning correlations on categories, we conducted a simple experiment. At first we trained the first layer with training data and prepared two datasets different from the training data. Then we produced two datasets by attaching the probabilistic outputs from the first layer to the inputs of the prepared datasets. After that we investigated the relationships between categories by training the rule-based classifier we had developed [11] (The rule-based classifier is based on successive, greedy generation of rules each of which covers most of the data at each time). We could find some relationships between the rules generated. Table 2 displays examples of discovered rules.

As we see from Table 2, there are many (strong and weak) relationships among categories. Generally, slightly dark images (such as the images with the possibility of $0.5 < \text{Night} < 0.6625$) are likely to be classified as night images without considering the correlations. But we can say these images are taken at daytime when they have a weak likelihood of Sunrise/Sunset.

We conclude that if we can build the classifiers reflecting many constraints well, the performance of categorization will improve significantly. In our experiments, the prediction results of these rule-based classifiers were not always good because many relationships had non-linear characteristics. As a remedy for this, we introduce additional layer of SVMs. The SVMs in the 2nd layer take the probabilities from the previous layer as inputs and outputs a new set of probabilities considering such hidden correlations among the categories.

3.3 Outputs of SVM Classifiers

Generally an SVM performs binary classification so it outputs only one of the two predefined numbers (for classes or categories). Also, an SVM does not output likelihoods in real values. So it is not possible to build the classification structure mentioned in this section by using standard SVMs.

Fortunately, there is a way to estimate likelihoods of outputs easily. Yang *et al.* used the confidence values as likelihoods of categorization results [3]. They assumed that the bigger confidence value means the stronger connection with the concept. They defined the confidence value as the distance of the input feature vector from the trained hyperplane of the SVM.

However this approach is not feasible because the meaning of distance varies according to the distribution of sample data. Let us assume that there are two different categories, A and B . For category A , positive samples are far from negative samples. But for category B , the distance between positive and negative samples is small. Then, the same confidence values of A and B

Table 3: Defined Categories and Distributions of Data (Unit: The Number of Samples in Each Category).

Three W's	Category	Brief Description	TD	Q1	Q2	VD
What	Waterside	River, sea or lake	397	103	142	43
	Snow	Snowcapped sites	419	92	101	38
	Self-portrait	Focus on a face (the most part)	374	111	101	35
	Food	Focus on food	400	90	97	40
	People	Many people	422	101	131	44
	Sunrise/Sunset	Sun or a glowing sky	404	94	91	40
	Unknown	No conspicuous object	-	550	440	218
When	Night	Night or in the dark	501	84	109	55
	Evening	Sun or dusk falling	479	93	90	50
	Day	In the bright light	504	964	904	353
Where	Stadium	Park, field or stand	300	109	108	41
	City	Buildings or roads	301	89	152	39
	Landscape	Mountain, river, sea or snowy sites	400	342	346	131
	Unknown	No conspicuous object	-	601	497	247

have different meaning, and it is obvious that the confidence value of B is more important. In addition, in our photo categorization approach, different feature subsets and kernels are used for each category. So the meanings of confidence values on categories can be different even if A and B have the same distribution. Therefore a different way of estimating likelihoods is needed and it must be available to compare the outputs without paying attention to the meanings of likelihoods.

Wu, Lin and Weng developed approaches for obtaining class probabilities in addition to the classification results of binary and multiclass classifiers [12] implemented in the LIBSVM software [13]. Unlike the confidence values of Yang *et al.*, the probabilistic outputs of the Wu's method all have the same ranges of 0 to 1, so the values can be compared with each other. We adopted the method to estimate the probabilities of SVM outputs.

4 Experiments

4.1 Data and Categories

First, we prepared four distinct chunks of home photos by using the feature extraction methods described in Section 2: training data (TD), test data 1 (Q1), test data 2 (Q2) for SVMs in the first layer, and validation data for deciding kernels of SVMs in the second layer (VD). All photos were collected by requests to authors' acquaintances or by downloading from personal blogs.

Home photo can take a variety of forms and be assigned to diverse categories. We focused on the approach of the five W's and one H, which is a formula for getting the full description of a situation. The system was unable to recognize "who" in an image. But it could detect whether a person was in it or not. Thus, the person could be regarded as an object of "what". Also, we cannot easily figure out why or how the picture was taken. So we finally considered only three W's which are "what", "when" and "where" as the highest groupings, and then defined categories (and descriptions or representative objects in the categories) of our interest under such groupings. Table 3 shows the detailed information on the data.

For the training dataset (TD), we gathered representative photos separately for each category.

For example, there are only the objects related to the stadium in stadium images regardless of when they were taken. The images of each category are the inputs of each SVM in the first layer. There are 4,901 photos in total in TD.

There are 1,141, 1,103 and 458 photos in total in Q1, Q2 and VD datasets, respectively. In Q1, Q2 and VD, all pictures belong to three categories: “*what*”, “*when*” and “*where*”. For instance, a photo can belong to *where-unknown*, *when-day*, and *what-food* but cannot belong to *where-city*, *what-snow* and *what-people*. In the case of “*what*”, if there are several objects in an image, the largest and centered object stands for the image. If there is no conspicuous object in the image, it is regarded as *unknown*. Figure 3 shows some sample images corresponding to the categories.

4.2 Experimental Process and Results

Previously we discussed the categorization process of home photos. For the practical applications of this, we need to go through the procedures finding the best classifiers in the first and second layers as explained in Section 3.

We construct two classification models, Model 1 and 2, for the verification of our system. For Model 1, we train SVMs in the first layer with TD, by choosing the feature subset and the kernel for all possible cases. We then discover the best classifier settings for all SVMs by selecting classifiers which yield the best classification results on the test data Q2 for each category. There are 12 categories excluding the unknown category, so 12 SVMs are trained in the first layer. Next, we train SVMs in the second layer with the classification result of Q2 (likelihoods) by the SVMs in the first layer. After that, in the same way as selecting the best classifiers in the first layer, we classify VD and decide kernels for each SVM that yields the best classification result. Similarly, we can construct Model 2 using TD and Q1 as training data.

Finally, we use Q1 as test data for Model 1 and Q2 as test data for Model 2 and evaluate the classification performance. The reason for using four different sets of data (i.e., TD, Q1, Q2, VD) is to derive robust classifiers with good generalization capability by considering data prepared at different times and/or by different people. Figure 4 displays the scheme of the photo categorization system and the learning procedure.

In the studies of image categorization, precision and recall are both important. Precision estimates how well the system removes what users do not want. Recall estimates how well it finds what users want. So we use the F1 measure (1) that combines precision and recall with an equal weight, as a performance criterion in finding the best classifiers.

$$F1 = 2 \cdot \textit{precision} \cdot \textit{recall} / (\textit{precision} + \textit{recall}). \quad (1)$$

The experimental results (excluding the unknown category) of Model 1 are shown in Table 4 and 5, with the selected parameters of each SVM and the performance on Q1 in terms of precision and recall. The results of Model 2 are shown in Table 6 and 7, with performance on Q2 (*Prec* means precision).

Figure 5(a), 5(b), 6(a) and 6(b) are the visualized results (in precision and recall) of Table 4, 5, 6 and 7, respectively.

We can see a great improvement in precision by using the second layer. The average precision was increased from 0.712 to 0.809 in Model 1 and 0.726 to 0.829 in Model 2. The city and the waterside categories are rather inaccurate, because the objects of a city such as buildings and roads are similar to artificial or indoor objects making it difficult to distinguish them, and the waterside photos have various forms of color, composition, shape, texture, and so on. But these difficulties are overcome by introducing the second layer of SVMs.











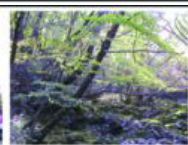


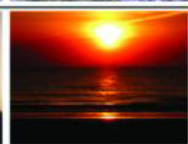



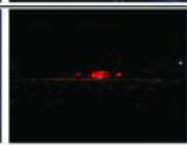







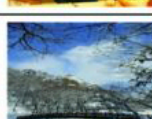
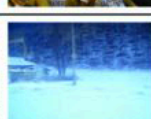



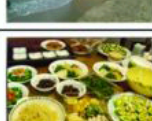





Where	City	Medium or long distance views of buildings or roads or both			
	Landscape	Medium or long distance views of mountains, woods, rivers, a sea, or a snowy landscape			
	Stadium	Photos of a baseball park or a football/soccer field focusing on a field or stands			
When	Day	Photos taken at daylight or in the bright light			
	Evening	Photos of rising or setting sun, or dusk falling			
	Night	Photos taken at night or in the dark			
What	Self-portrait	Photos of focusing on a face (the most part of a photo is a face)			
	People	Portrait photos (just enough to distinguish people's faces)			
	Snow	Various photos of snowcapped places			
	Waterside	Photos including scenes of rivers, a sea, or lakes			
	Food	Photos that the mainly focused object is food			
	Sunrise/Sunset	Photos of rising or setting sun, or a glow in the sky			

Figure 3: Categories of Home Photos and Their Descriptions Including Sample Photographs.

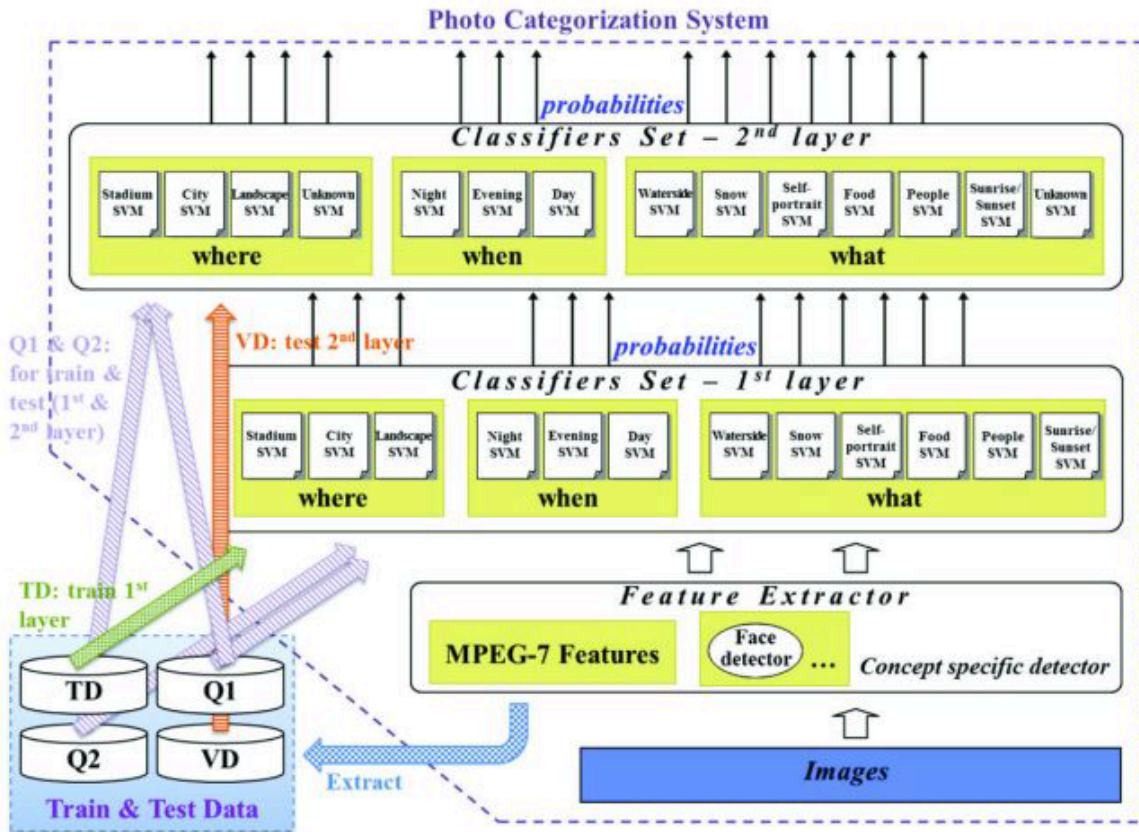


Figure 4: Training Classifiers and Validation with Datasets in Table 3.

Table 4: Selected Settings and Performances of Model 1 (First Layer).

W's	Category	Selected Features	Kernel	Prec	Recall	F1
What	Waterside	All features	RBF	0.446	0.563	0.497
	Snow	All features	Poly	0.586	0.771	0.666
	Self-portrait	DCD, FD	RBF	0.919	0.819	0.866
	Food	DCD, EHD, FD	Poly	0.804	0.733	0.767
	People	FD	RBF	0.969	0.623	0.759
	Sunrise/Sunset	CSD, CLD, EHD	Poly	0.834	0.914	0.873
When	Night	DCD, EHD, FD	Poly	0.650	0.797	0.716
	Evening	DCD, CLD, EHD	Poly	0.750	0.870	0.805
	Day	DCD, CLD, EHD	Poly	0.985	0.845	0.910
Where	Stadium	CSD, EHD, FD	Poly	0.707	0.908	0.795
	City	DCD, CLD, EHD, FD	RBF	0.331	0.674	0.444
	Landscape	CSD, DCD, CLD, EHD	Poly	0.568	0.751	0.647
Average				0.712	0.772	0.729

Yang *et al.* defined categories similar or identical to ours (i.e., terrain (corresponding to landscape), night-scene (night), snowspace (snow), sunset (sunrise/set), and waterside) [3]. Even though the comparison with the results reported in the paper is not perfect, we see our system produces higher precision and lower recall in general, and is significantly better in snow and sunrise/set categories in particular. The comparison result of our approach (the second layer of

Table 5: Selected Settings and Performances of Model 1 (Second Layer).

W's	Category	Kernel	Prec	Recall	F1
What	Waterside	RBF	0.584	0.300	0.397
	Snow	RBF	0.797	0.641	0.710
	Self-portrait	RBF	0.918	0.810	0.861
	Food	RBF	0.905	0.533	0.671
	People	Poly	0.969	0.623	0.759
	Sunrise/Sunset	RBF	0.912	0.882	0.897
When	Night	Poly	0.744	0.761	0.752
	Evening	Poly	0.951	0.838	0.891
	Day	Poly	0.978	0.973	0.975
Where	Stadium	Sigmoid	0.725	0.944	0.820
	City	Poly	0.616	0.595	0.605
	Landscape	RBF	0.606	0.567	0.586
Average			0.809	0.706	0.744

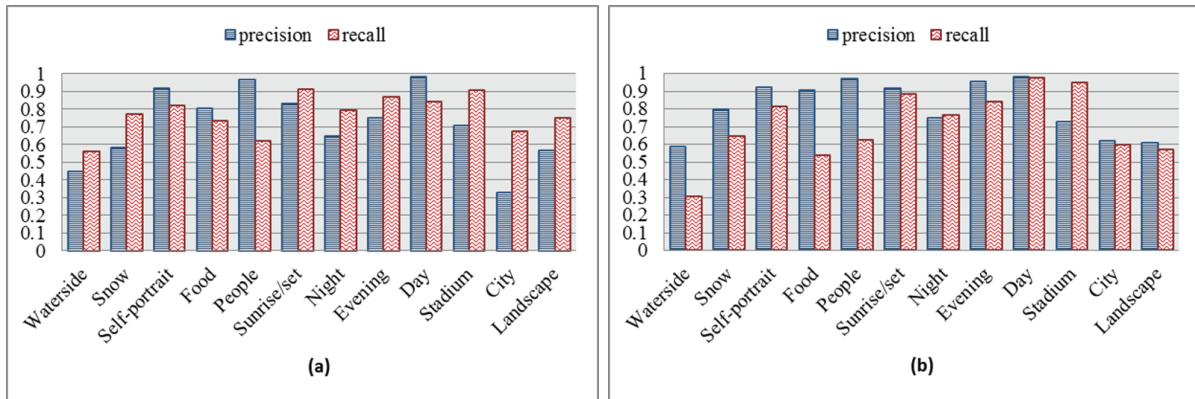


Figure 5: Classification Performance in Model 1 ((a): First Layer, (b): Second Layer).

Table 6: Selected Settings and Performances of Model 2 (First Layer).

W's	Category	Selected Features	Kernel	Prec	Recall	F1
What	Waterside	CSD, DCD, CLD, EHD	Poly	0.494	0.612	0.547
	Snow	All features	Poly	0.758	0.900	0.823
	Self-portrait	CLD, FD	Poly	0.814	0.435	0.567
	Food	DCD, CLD, EHD, FD	RBF	0.733	0.793	0.762
	People	FD	RBF	0.777	0.427	0.567
	Sunrise/Sunset	CSD, DCD, CLD, EHD	Poly	0.802	0.714	0.755
When	Night	DCD	RBF	0.862	0.807	0.834
	Evening	CSD, DCD, CLD, EHD	Poly	0.634	0.733	0.680
	Day	CLD, EHD	Poly	0.980	0.825	0.896
Where	Stadium	CSD, CLD, EHD, FD	Poly	0.701	0.675	0.688
	City	DCD, CLD, EHD, FD	RBF	0.492	0.651	0.560
	Landscape	CSD, DCD, CLD, EHD	Poly	0.659	0.841	0.739
Average				0.726	0.701	0.702

Table 7: Selected Settings and Performances of Model 2 (Second Layer).

W's	Category	Kernel	Prec	Recall	F1
What	Waterside	Poly	0.563	0.373	0.449
	Snow	RBF	0.890	0.801	0.843
	Self-portrait	Poly	0.854	0.405	0.550
	Food	Sigmoid	0.745	0.783	0.763
	People	Sigmoid	0.730	0.351	0.474
	Sunrise/Sunset	Sigmoid	0.869	0.659	0.750
When	Night	RBF	0.926	0.577	0.711
	Evening	Sigmoid	0.904	0.633	0.745
	Day	RBF	0.931	0.991	0.960
Where	Stadium	Poly	0.850	0.472	0.607
	City	Poly	0.905	0.315	0.468
	Landscape	Sigmoid	0.775	0.699	0.735
Average			0.829	0.588	0.671

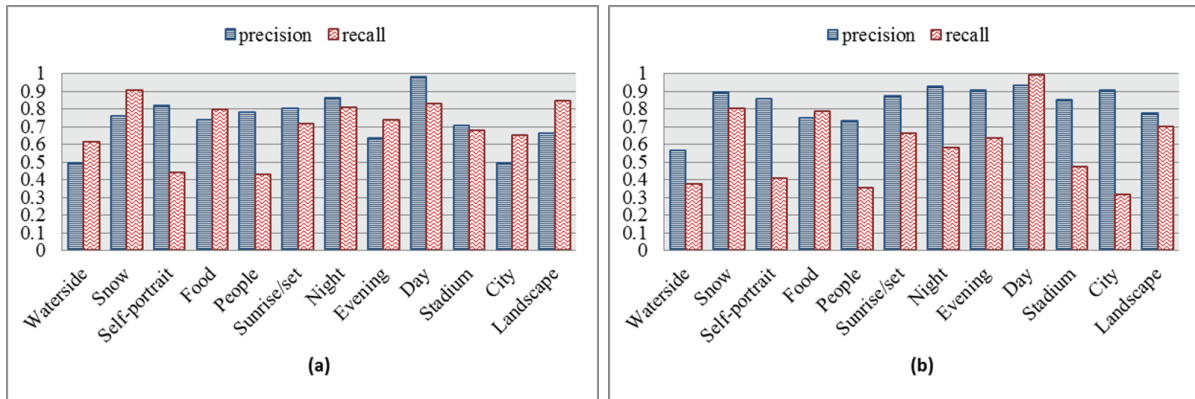


Figure 6: Classification Performance in Model 2 ((a): First Layer, (b): Second Layer).

Model 1) and Yang *et al.*'s method is shown in Figure 7.

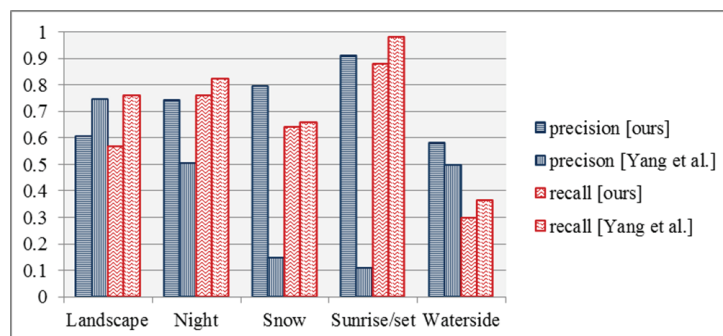


Figure 7: Performance Comparison between the Method of [3] and Our Method.

Moreover, there is no evaluation of computation time in [3]. We surmise their method would be much slower than ours since they use time-consuming features like the Homogeneous Texture Descriptor, and extract such features quite often for every five local regions. In contrast, our algorithm uses fast feature extraction methods and does not extract features repeatedly from an

image. Actually, it takes about less than one second for categorizing a photo on the Samsung Omnia 2 Smartphone (ARM11 800Mhz CPU). So our algorithm is suitable for use on handheld devices.

Figure 8 displays snapshots of sample runs of the home photo categorization and browsing software performed on the actual device. Users can browse photos by selecting “*what*”, “*when*” and “*where*” categories. Also the software supports the auto-categorizing function for photographing by the built-in camera.

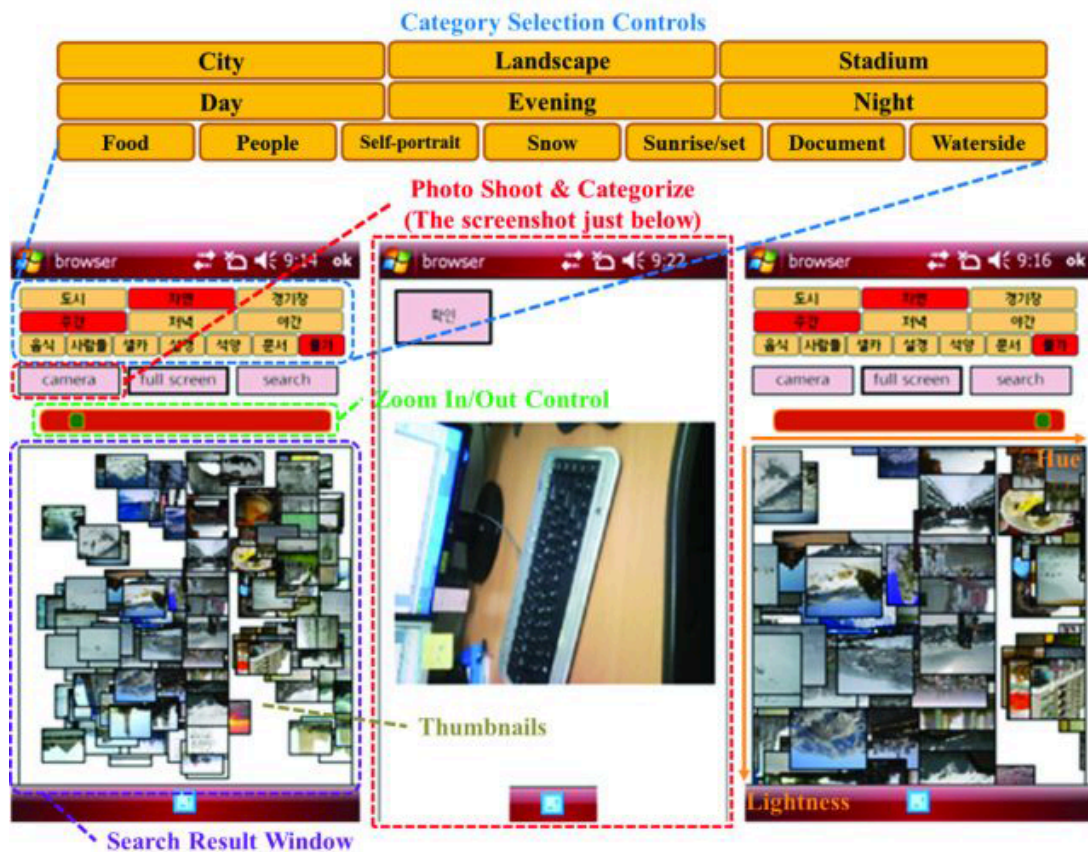


Figure 8: Sample Runs of The Home Photo Categorization and Browsing Software.

5 Conclusion

In this paper, we proposed an efficient home photo categorization method using fast MPEG-7 descriptors and the face extractor, and a two-layered classifiers of SVM. The classifiers in the first layer are trained to assign images into predefined categories, and the ones in the second layer attempt to improve the classification performance by considering the relationships and constraints among the categories. In the way to construct the multi-level classifiers, we also considered feature and kernel selections and obtained the best feature subsets and kernel functions. Our method was compared with one of the home photo categorization methods and verified to produce outstanding performance with less computational overhead, which is a prerequisite for the implementation in real handheld devices.

In spite of the effectiveness of the proposed method, there are several challenging issues. First, as the face feature is the most important factor in distinguishing people and self-portrait photos

from others, we may as well implement new feature extractors specialized in extracting unique features of photos in certain categories (e.g. city, landscape). The extractors should have low computational cost in order to support real-time categorization. By applying the state-of-the-art technology like the fast object detection method [14], we may be able to obtain better results. We are currently developing extracting tools for additional object-based features (e.g., buildings for city category) to enhance our categorization system. As far as face detection, the current algorithm works only with frontal faces, which can be extended to consider rotated objects as proposed in [15]. Also, the recall of the second layer's outputs was not improved with the concept of relational learning, unlike precision, so we need to compensate for this weak point to make our method more powerful. In addition, finding the best parameter settings for SVM is of significance instead of blindly relying on widely used ones.

Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology to Jihoon Yang, the corresponding author (2013R1A1A2012502), and by the IT R&D program of MSIP/KEIT (10044615: Development of Open-Platform/Social Media Production and Delivery System for Fused Creation, Editing, and Playing of Broadcasting Media Contents on Cloud Environments).

Bibliography

- [1] J.H. Lim, J.S. Jin, Unifying local and global content-based similarities for home photo retrieval, *Proceedings of 2004 International Conference on Image Processing*, 4:2371-2374, 2004.
- [2] Y. Chen and J.Z. Yang, Image Categorization by Learning and Reasoning with Regions, *The Journal of Machine Learning Research*, 5:913-939, 2004.
- [3] S.J. Yang, S.K. Kim, K.S. Seo, Y.M. Ro, J.Y. Kim, Y.S. Seo, Semantic categorization of digital home photo using photographic region templates, *Proceedings of 2005 Information retrieval research in Asia*, 43(2):503-514, 2007.
- [4] J.M. Martínez, MPEG-7 Overview, *ISO/IEC JTC1/SC29/WG11N6828*, 2004.
- [5] C.J.C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, *Data Mining and Knowledge Discovery*, 2(2):121-167, 1998.
- [6] T. Mitchell, *Machine Learning*, McGraw Hill, 1998.
- [7] H. Eidenberger, Statistical analysis of content-based MPEG-7 descriptors for image retrieval, *Multimedia Systems*, 10:84-97, 2004.
- [8] J.S. Yu, J.H. Nang, An Optimization Method for Extraction of MPEG-7 Color Structure Descriptor and Dominant Color Descriptor, *Proceedings of Korea Computer Congress 2009*, 36(1A):320-321, 2009.
- [9] Institute for Integrated Circuits, Technische Universit Munchen, *MPEG-7 XM Software*, Germany, 2003. Available (Online):
http://standards.iso.org/ittf/PubliclyAvailableStandards/c035364_ISO_IEC_15938-6%28E%29_Reference_Software.zip

- [10] B. Fröba, A. Ernst, Face Detection with the Modified Census Transform, *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, 0:91-96, 2004.
- [11] B.H. Oh, J.H. Yang, Discovering Classification Rules using Genetic Algorithm, *Proceedings of Korea Computer Congress 2009*, 36(1C):480-485, 2009.
- [12] T.F. Wu, C.J. Lin, R.C. Weng, Probability Estimates for Multi-class Classification by Pairwise Coupling, *Journal of Machine Learning Research*, 5:975-1005, 2003.
- [13] C.C. Chang, C.J. Lin, *LIBSVM : a library for support vector machines*, 2001.
- [14] M.M. Jlasi, A. Douik, H. Messaoud, Objects Detection by Singular Value Decomposition Technique in Hybrid Color Space: Application to Football Images, *International Journal of Computers Communications & Control*, 5(2):193-204, 2010.
- [15] T. Barbu, An Automatic Face Detection System for RGB Images, *International Journal of Computers Communications & Control*, 6(1):21-32, 2011.

Feedback Linearization with Fuzzy Compensation for Uncertain Nonlinear Systems

M.C. Tanaka, J.M.M. Fernandes, W.M. Bessa

**Marcelo C. Tanaka, Josiane M.M. Fernandes,
Wallace M. Bessa**

Departamento de Engenharia Mecânica,
Universidade Federal do Rio Grande do Norte
Campus Universitário Lagoa Nova, Natal,
RN 59072-970, Brazil
marcelotanaka.eng.mec@gmail.com,
josiane.eng.mec@gmail.com, wmbessa@ct.ufrn.br

Abstract:

This paper presents a nonlinear controller for uncertain single-input–single-output (SISO) nonlinear systems. The adopted approach is based on the feedback linearization strategy and enhanced by a fuzzy inference algorithm to cope with modeling inaccuracies and external disturbances that can arise. The boundedness and convergence properties of the tracking error vector are analytically proven. An application of the proposed control scheme to a second-order nonlinear system is also presented. The obtained numerical results demonstrate the improved control system performance.

Keywords: feedback linearization, fuzzy logic, nonlinear control, Van der Pol oscillator.

1 Introduction

Due to its simplicity, feedback linearization scheme is commonly applied in industrial control systems, specially in the field of industrial robotics. The main idea behind this control method is the development of a control law that allows the transformation of the original dynamical system into an equivalent but simpler one [11]. Although feedback linearization represents a very simple approach, an important handicap is the requirement of a perfectly known dynamical system, in order to ensure the exponential convergence of the tracking error.

On this basis, much effort has been made to combine feedback linearization with intelligent algorithms in order to improve the trajectory tracking of uncertain nonlinear systems. The most common strategies are based on artificial neural networks [2, 4, 9, 10, 13] or fuzzy logic [1, 3, 5, 6]. A drawback of these approaches is that both neural networks or fuzzy logic are used to model the entire plant, which means that a large computational effort is normally required to characterize system dynamics.

Considering that the designer of the control system usually has at least some knowledge of the plant to be controlled, a nonlinear controller is proposed in this paper to compensate for the uncertainties of single-input-single-output (SISO) nonlinear systems. The adopted approach is based on the feedback linearization method, but enhanced by a fuzzy inference system to cope with modeling imprecisions and external disturbances that can arise. This approach requires a reduced number of fuzzy sets and rules and consequently simplifies the design process. The boundedness and convergence properties of the closed-loop signals are analytically proven and numerical simulations are carried out in order to demonstrate the improved performance of the proposed control scheme.

2 Feedback Linearization

Consider a class of n^{th} -order nonlinear systems:

$$x^{(n)} = f(\mathbf{x}, t) + b(\mathbf{x}, t)u + d \quad (1)$$

where u is the control input, the scalar variable x is the output of interest, $x^{(n)}$ is the n -th time derivative of x , $\mathbf{x} = [x, \dot{x}, \dots, x^{(n-1)}]$ is the system state vector, $f, b : \mathbb{R}^n \rightarrow \mathbb{R}$ are both nonlinear functions and d is assumed to represent all uncertainties and unmodeled dynamics regarding system dynamics, as well as any external disturbance that can arise.

In respect of the disturbance-like term d , the following assumption will be made:

Assumption 1. *The disturbance d is unknown but continuous and bounded, i. e. $|d| \leq \delta$.*

Let us now define an appropriate control law based on conventional feedback linearization scheme that ensures the tracking of a desired trajectory $\mathbf{x}_d = [x_d, \dot{x}_d, \dots, x_d^{(n-1)}]$, i. e. the controller should assure that $\tilde{\mathbf{x}} \rightarrow 0$ as $t \rightarrow \infty$, where $\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}_d = [\tilde{x}, \dot{\tilde{x}}, \dots, \tilde{x}^{(n-1)}]$ is the related tracking error.

On this basis, assuming that the state vector \mathbf{x} is available to be measured and system dynamics is perfectly known, i. e. there is no modeling imprecision nor external disturbance ($d = 0$) and the functions f and b are well known, with $|b(\mathbf{x}, t)| > 0$, the following control law:

$$u = b^{-1}(-f + x_d^{(n)} - k_0\tilde{x} - k_1\dot{\tilde{x}} - \dots - k_{n-1}\tilde{x}^{(n-1)}) \quad (2)$$

guarantees that $\mathbf{x} \rightarrow \mathbf{x}_d$ as $t \rightarrow \infty$, if the coefficients k_i ($i = 0, 2, \dots, n-1$) make the polynomial $p^n + k_{n-1}p^{n-1} + \dots + k_0$ a Hurwitz polynomial [11].

The convergence of the closed-loop system could be easily established by substituting the control law (2) in the nonlinear system (1). The resulting dynamical system could be rewritten by means of the tracking error:

$$\tilde{x}^{(n)} + k_{n-1}\tilde{x}^{(n-1)} + \dots + k_1\dot{\tilde{x}} + k_0\tilde{x} = 0 \quad (3)$$

where the related characteristic polynomial is Hurwitz.

However, since in real-world applications the nonlinear system (1) is often not perfectly known, the control law (2) based on conventional feedback linearization is not sufficient to ensure the exponential convergence of the tracking error to zero.

Thus, we propose the adoption of fuzzy inference system within the control law, in order to compensate for d and to enhance the feedback linearization controller.

3 Fuzzy Inference System

Because of the possibility to express human experience in an algorithmic manner, fuzzy logic has been largely employed in the last decades to both control and identification of dynamical systems.

The adopted fuzzy inference system is the zero order TSK (Takagi–Sugeno–Kang), with the r^{th} rule stated in a linguistic manner as follows:

$$\text{If } \tilde{x} \text{ is } \tilde{X}_r, \dot{\tilde{x}} \text{ is } \dot{\tilde{X}}_r, \dots, \text{ and } \tilde{x}^{(n-1)} \text{ is } \tilde{X}_r^{(n-1)}, \text{ then } \hat{d}_r = \hat{D}_r \quad ; \quad r = 1, 2, \dots, N$$

where $\tilde{X}_r, \dot{\tilde{X}}_r, \dots$, and $\tilde{X}_r^{(n-1)}$ are fuzzy sets, whose membership functions could be properly chosen, and \hat{D}_r is the output value of each one of the N fuzzy rules.

Considering that each rule defines a numerical value as output \hat{D}_r , the final output \hat{d} can be computed by a weighted average:

$$\hat{d}(\tilde{\mathbf{x}}) = \frac{\sum_{r=1}^N w_r \cdot \hat{D}_r}{\sum_{r=1}^N w_r} \quad (4)$$

or, similarly,

$$\hat{d}(\tilde{\mathbf{x}}) = \hat{\mathbf{D}}^T \Psi(\tilde{\mathbf{x}}) \quad (5)$$

where, $\hat{\mathbf{D}} = [\hat{D}_1, \hat{D}_2, \dots, \hat{D}_N]$ is the vector containing the attributed values \hat{D}_r to each rule r , $\Psi(\tilde{\mathbf{x}}) = [\psi_1, \psi_2, \dots, \psi_N]$ is a vector with components $\psi_r(\tilde{\mathbf{x}}) = w_r / \sum_{r=1}^N w_r$ and w_r is the firing strength of each rule, which can be computed from the membership values with any fuzzy intersection operator (t-norm).

4 Fuzzy Feedback Linearization

Considering that fuzzy logic can perform universal approximation [7], we propose the adoption of a TSK fuzzy inference system within the feedback linearization controller to compensate for modeling inaccuracies and consequently enhance the trajectory tracking of uncertain nonlinear systems.

Therefore, the control law with the fuzzy compensation scheme can be stated as follows

$$u = b^{-1}[-f + x_d^{(n)} - k_0 \tilde{x} - k_1 \dot{\tilde{x}} - \dots - k_{n-1} \tilde{x}^{(n-1)} - \hat{d}(\tilde{\mathbf{x}})] \quad (6)$$

and the related closed-loop system is:

$$\tilde{x}^{(n)} + k_{n-1} \tilde{x}^{(n-1)} + \dots + k_1 \dot{\tilde{x}} + k_0 \tilde{x} = \tilde{d} \quad (7)$$

with $\tilde{d} = \hat{d} - d$.

Now, defining $\mathbf{k}^T \tilde{\mathbf{x}} = k_{n-1} \tilde{x}^{(n-1)} + \dots + k_1 \dot{\tilde{x}} + k_0 \tilde{x}$, where $\mathbf{k} = [c_0 \lambda^n, c_1 \lambda^{n-1}, \dots, c_{n-1} \lambda]$, λ is a strictly positive constant and c_i states for binomial coefficients, *i. e.*

$$c_i = \binom{n}{i} = \frac{n!}{(n-i)! i!}, \quad i = 0, 1, \dots, n-1 \quad (8)$$

the convergence of the closed-loop signals to a bounded region is assured.

Theorem 2. Consider the uncertain nonlinear system (1) and Assumption 1, then the fuzzy feedback linearization controller defined by (5) and (6) ensures the exponential convergence of the tracking error vector to a closed region $\Omega = \{\mathbf{x} \in \mathbb{R}^n \mid |\tilde{x}^{(i)}| \leq \zeta_i \lambda^{i-n} \varepsilon, i = 0, 1, \dots, n-1\}$, with ζ_i defined by (9).

$$\zeta_i = \begin{cases} 1 & \text{for } i = 0 \\ 1 + \sum_{j=0}^{i-1} \binom{i}{j} \zeta_j & \text{for } i = 1, 2, \dots, n-1. \end{cases} \quad (9)$$

Proof: Considering the universal approximation feature of fuzzy logic [7], the output of the adopted inference system (5) can approximate the disturbance d to an arbitrary degree of accuracy, *i. e.* $|\hat{d}(\tilde{x}) - d| \leq \varepsilon$ for an arbitrary $\varepsilon > 0$. Thus, from (7) one has

$$|\tilde{x}^{(n)} + k_{n-1} \tilde{x}^{(n-1)} + \dots + k_1 \dot{\tilde{x}} + k_0 \tilde{x}| \leq \varepsilon \quad (10)$$

From (8), inequality (10) may be rewritten as

$$-\varepsilon \leq \tilde{x}^{(n)} + c_{n-1} \lambda \tilde{x}^{(n-1)} + \dots + c_1 \lambda^{n-1} \dot{\tilde{x}} + c_0 \lambda^n \tilde{x} \leq \varepsilon \quad (11)$$

Multiplying (11) by $e^{\lambda t}$ yields

$$-\varepsilon e^{\lambda t} \leq \frac{d^n}{dt^n}(\tilde{x}e^{\lambda t}) \leq \varepsilon e^{\lambda t} \quad (12)$$

Integrating (12) between 0 and t gives

$$-\frac{\varepsilon}{\lambda}e^{\lambda t} + \frac{\varepsilon}{\lambda} \leq \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) - \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t})\Big|_{t=0} \leq \frac{\varepsilon}{\lambda}e^{\lambda t} - \frac{\varepsilon}{\lambda} \quad (13)$$

or conveniently rewritten as

$$-\frac{\varepsilon}{\lambda}e^{\lambda t} - \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \leq \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \leq \frac{\varepsilon}{\lambda}e^{\lambda t} + \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \quad (14)$$

The same reasoning can be repeatedly applied until the n^{th} integral of (12) is reached:

$$\begin{aligned} -\frac{\varepsilon}{\lambda^n}e^{\lambda t} - \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \frac{t^{n-1}}{(n-1)!} - \dots + \\ - \left(|\tilde{x}(0)| + \frac{\varepsilon}{\lambda^n} \right) \leq \tilde{x}e^{\lambda t} \leq \frac{\varepsilon}{\lambda^n}e^{\lambda t} + \\ + \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \frac{t^{n-1}}{(n-1)!} + \dots + \left(|\tilde{x}(0)| + \frac{\varepsilon}{\lambda^n} \right) \end{aligned} \quad (15)$$

Furthermore, dividing (15) by $e^{\lambda t}$, it can be easily verified that, for $t \rightarrow \infty$,

$$-\frac{\varepsilon}{\lambda^n} \leq \tilde{x}(t) \leq \frac{\varepsilon}{\lambda^n} \quad (16)$$

Considering the $(n-1)^{\text{th}}$ integral of (12)

$$\begin{aligned} -\frac{\varepsilon}{\lambda^{n-1}}e^{\lambda t} - \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \frac{t^{n-2}}{(n-2)!} - \dots + \\ - \left(|\dot{\tilde{x}}(0)| + \frac{\varepsilon}{\lambda^{n-1}} \right) \leq \frac{d}{dt}(\tilde{x}e^{\lambda t}) \leq \frac{\varepsilon}{\lambda^{n-1}}e^{\lambda t} + \\ + \left(\left| \frac{d^{n-1}}{dt^{n-1}}(\tilde{x}e^{\lambda t}) \right|_{t=0} + \frac{\varepsilon}{\lambda} \right) \frac{t^{n-2}}{(n-2)!} + \dots + \left(|\dot{\tilde{x}}(0)| + \frac{\varepsilon}{\lambda^{n-1}} \right) \end{aligned} \quad (17)$$

and noting that $d(\tilde{x}e^{\lambda t})/dt = \dot{\tilde{x}}e^{\lambda t} + \tilde{x}\lambda e^{\lambda t}$, by imposing the bounds (16) to (17) and dividing again by $e^{\lambda t}$, it follows that, for $t \rightarrow \infty$,

$$-2\frac{\varepsilon}{\lambda^{n-1}} \leq \dot{\tilde{x}}(t) \leq 2\frac{\varepsilon}{\lambda^{n-1}} \quad (18)$$

Now, applying the bounds (16) and (18) to the $(n-2)^{\text{th}}$ integral of (12) and dividing once again by $e^{\lambda t}$, it follows that, for $t \rightarrow \infty$,

$$-6\frac{\varepsilon}{\lambda^{n-2}} \leq \ddot{\tilde{x}}(t) \leq 6\frac{\varepsilon}{\lambda^{n-2}} \quad (19)$$

The same procedure can be successively repeated until the bounds for $\tilde{x}^{(n-1)}$ are achieved:

$$-\left[1 + \sum_{i=0}^{n-2} \binom{n-1}{i} \zeta_i\right] \frac{\varepsilon}{\lambda} \leq \tilde{x}^{(n-1)} \leq \left[1 + \sum_{i=0}^{n-2} \binom{n-1}{i} \zeta_i\right] \frac{\varepsilon}{\lambda} \quad (20)$$

where the coefficients ζ_i ($i = 0, 1, \dots, n-2$) are related to the previously obtained bounds of each $\tilde{x}^{(i)}$ and can be summarized as in (9).

In this way, by inspection of the integrals of (12), as well as (16), (18), (19), (20) and the other omitted bounds, it follows that the tracking error exponentially converges to the n -dimensional box determined by the limits $|\tilde{x}^{(i)}| \leq \zeta_i \lambda^{i-n} \varepsilon$, $i = 0, 1, \dots, n-1$, where ζ_i is defined by (9). \square

Corollary 3. *It must be noted that the proposed control scheme provides a smaller tracking error when compared with the conventional feedback linearization controller. By setting the output of the fuzzy inference system to zero, $\hat{d}(\tilde{x}) = 0$, Theorem 2 implies that the resulting bounds are $|\tilde{x}^{(i)}| \leq \zeta_i \lambda^{i-n} \delta$, $i = 0, 1, \dots, n-1$. Considering that $\varepsilon < \delta$, from the universal approximation feature of \hat{d} , it can be concluded that the tracking error obtained with the fuzzy feedback linearization controller is smaller than the associated with the conventional scheme.*

5 Illustrative Example

In order to illustrate the controller design methodology, consider a controlled Van der Pol oscillator

$$\ddot{x} - \mu(1 - x^2)\dot{x} + x = v \quad (21)$$

with a dead-zone in the control input defined according to

$$v = \begin{cases} u + 0.2 & \text{if } u \leq -0.2 \\ 0 & \text{if } -0.2 < u < 0.2 \\ u - 0.2 & \text{if } u \geq 0.2 \end{cases} \quad (22)$$

For control purposes, equation (22) can be rewritten as a combination of a linear and a saturation function [8, 12]:

$$v = u + d(u) \quad (23)$$

where $d(u)$ can be obtained from (22) and (23) as:

$$d(u) = \begin{cases} 0.2 & \text{if } u \leq -0.2 \\ -u & \text{if } -0.2 < u < 0.2 \\ -0.2 & \text{if } u \geq 0.2 \end{cases} \quad (24)$$

Based on (6) and considering $d(u)$ as uncertainty, a fuzzy feedback linearization controller can be chosen as follows

$$u = x - \mu(1 - x^2)\dot{x} + \ddot{x}_d - 2\lambda\dot{x} - \lambda^2\tilde{x} - \hat{d}(\tilde{x}, \dot{\tilde{x}}) \quad (25)$$

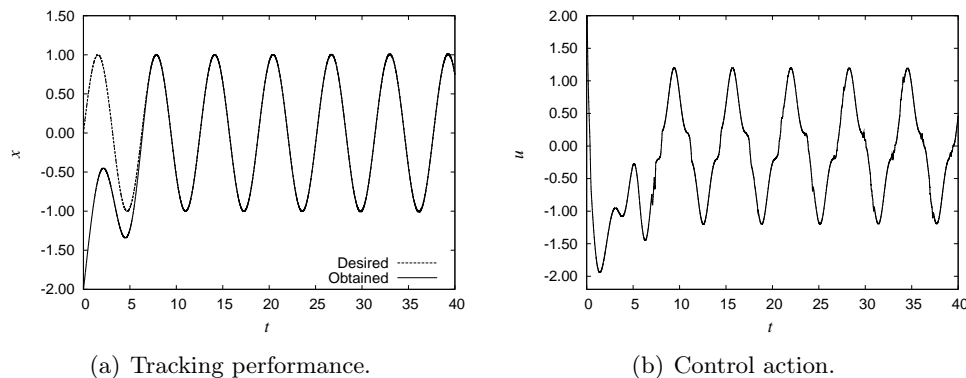
In order to evaluate the performance of the proposed control law (25), a numerical simulation was carried out. The simulation study was performed with an implementation in C, with sampling rates of 500 Hz for control system and 1 kHz for the Van der Pol oscillator, and the differential equations were numerically solved using the fourth order Runge-Kutta method. The chosen parameters for the Van der Pol oscillator and controller were $\mu = 1$ and $\lambda = 0.8$.

Regarding the fuzzy inference system, the number of fuzzy rules and the type of the membership functions, as well as how they are distributed over the input space, could be heuristically defined to accommodate designer's experience and experimental knowledge. The fuzzy rule base adopted in this work is presented in Table 1, where NB, NM, NS, ZO, PS, PM and PB represent, respectively, Negative-Big, Negative-Medium, Negative-Small, Zero, Positive-Small, Positive-Medium and Positive-Big. Triangular and trapezoidal (at the ends) membership functions are adopted for both \tilde{X}_r and $\dot{\tilde{X}}_r$, with the central values defined respectively as $C_{\tilde{x}} = \{-20; -2; -0.2; 0.0; 0.2; 2; 20\} \times 10^{-2}$ and $C_{\dot{\tilde{x}}} = \{-16; -1.6; -0.16; 0.0; 0.16; 1.6; 16\} \times 10^{-2}$. The chosen fuzzy intersection operator was the minimum t-norm. It should be also emphasized that the input space could be partitioned and represented in many other ways, and that the system designer may test each one of them in order to improve the output value \hat{d} . With respect to the output of each rule, the following values were heuristically adopted for NB to PB: $\hat{D}_r = \{-20; -5; -2.5; 0.0; 2.5; 5; 20\}$.

Table 1: Adopted fuzzy rule base.

$\tilde{x} / \dot{\tilde{x}}$	NB	NM	NS	ZO	PS	PM	PB
NB	PB	PB	PB	PM	PM	PS	ZO
NM	PB	PB	PM	PM	PS	ZO	NS
NS	PB	PM	PM	PS	ZO	NS	NM
ZO	PM	PM	PS	ZO	NS	NM	NM
PS	PM	PS	ZO	NS	NM	NM	NB
PM	PS	ZO	NS	NM	NM	NB	NB
PB	ZO	NS	NM	NM	NB	NB	NB

In this way, considering that the initial state and initial desired state are not equal, $\tilde{\mathbf{x}}(0) = [-2.0, -0.4]$, Figures 1–3 show the obtained results for the tracking of $\mathbf{x}_d = [\sin t, \cos t]$.


 Figure 1: Trajectory tracking with $\mathbf{x}_d = [\sin t, \cos t]$.

As observed in Figure. 1(a), even in the presence of modeling imprecisions, the proposed control scheme allows the actuated Van der Pol oscillator to track the desired trajectory.

Now, in order to demonstrate the improved performance of the fuzzy feedback linearization controller, the tracking error associated with the last simulation is shown in Fig. 2. For comparison purposes, the tracking error obtained with conventional feedback linearization is also presented. It can be easily verified that the proposed controller provides a smaller tracking error

when compared with the conventional one.

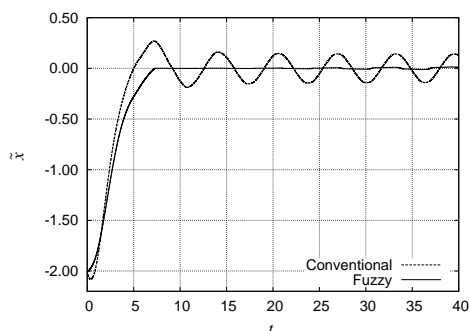


Figure 2: Tracking error with conventional and fuzzy feedback linearization.

The phase portraits of the tracking errors obtained with conventional as well as fuzzy feedback linearization are shown in Fig. 3. Note that the convergence region related to the proposed control scheme is much smaller than the associated with its uncompensated counterpart, which confirms Corollary 3.

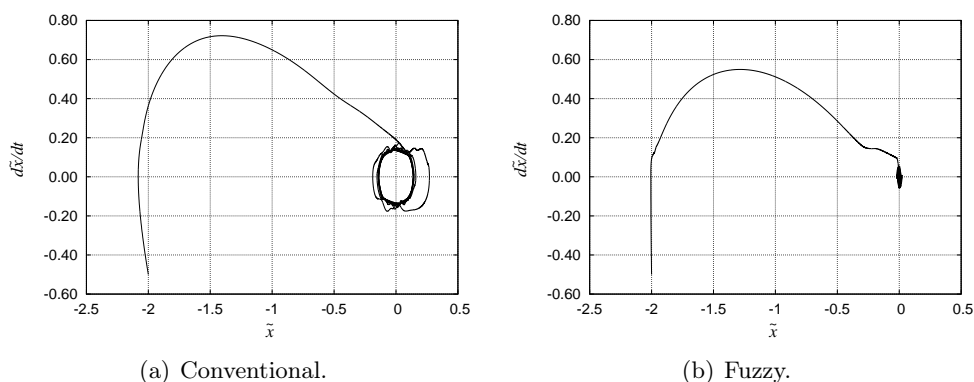


Figure 3: Phase portrait of the error with conventional and fuzzy feedback linearization.

6 Concluding Remarks

In this paper, a fuzzy feedback linearization controller is developed to deal with uncertain single-input–single-output nonlinear systems. To enhance the tracking performance, the feedback linearization controller is combined with a fuzzy inference system for uncertainty/disturbance compensation. The boundedness and convergence properties of the tracking error vector are analytically proven. To evaluate the control system performance, the proposed scheme is applied to the Van der Pol oscillator. By means of numerical simulations, the improved performance over the conventional feedback linearization controller is confirmed.

Acknowledgments

The authors would like to acknowledge the support of the Brazilian National Research Council (CNPq), the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES), the Brazilian National Agency of Petroleum, Natural Gas and Biofuels (ANP) and the German Academic Exchange Service (DAAD).

Bibliography

- [1] Boukezzoula, R.; Galichet, S.; Foulloy, L. (2007); Fuzzy feedback linearizing controller and its equivalence with the fuzzy nonlinear internal model control structure, *Int J Appl Math Comput Sci*, ISSN 1641-876X, 17(2):233-248.
- [2] Chen, F.C. (1990); Back-propagation neural networks for nonlinear self-tuning adaptive control, *IEEE Control Syst Mag*, ISSN 0272-1708, 10(3):44-48.
- [3] Couceiro, M.S.; Ferreira, N.M.F.; Machado, J.A.T. (2012); Hybrid adaptive control of a dragonfly model, *Commun Nonlinear Sci Numer Simul*, ISSN 1007-5704, 17(2):893-903.
- [4] Deng, H.; Li, H.X.; Wu, Y.H. (2008); Feedback-linearization-based neural adaptive control for unknown nonaffine nonlinear discrete-time systems, *IEEE Trans Neural Netw*, ISSN 1045-9227, 19(9):1615-1625.
- [5] Hojati, M.; Gazor, S. (2002); Hybrid adaptive fuzzy identification and control of nonlinear systems, *IEEE Trans Fuzzy Syst*, ISSN 1063-6706, 10(2):198-210.
- [6] Kang, H.J.; Kwon, C.; Lee, H.; Park, M. (1998); Robust stability analysis and design method for the fuzzy feedback linearization regulator, *IEEE Trans Fuzzy Syst*, ISSN 1063-6706, 6(4):464-472.
- [7] Kosko, B. (1994); Fuzzy systems as universal approximators, *IEEE Trans Comput*, 43(11):1329-1333.
- [8] Lewis, F.L.; Tim, W.K.; Wang, L.Z.; Li, Z.X. (1999); Deadzone compensation in motion control systems using adaptive fuzzy logic control, *IEEE Trans Control Syst Technol*, ISSN 1063-6536, 7(6):731-742.
- [9] Lu, Z.; Shieh, L.S.; Chen, G.; Coleman, N.P. (2006); Adaptive feedback linearization control of chaotic systems via recurrent high-order neural networks, *Inf Sci*, ISSN 0020-0255, 176(16):2337-2354.
- [10] Pedro J.O.; Dahunsi, O.A. (2011); Neural network based feedback linearization control of a servo-hydraulic vehicle suspension system, *Int J Appl Math Comput Sci*, ISSN 1641-876X, 21(1):137-147.
- [11] Slotine, J.J.E.; Li, W. (1991); *Applied Nonlinear Control*, Prentice Hall.
- [12] Wang, X.S.; Su, C.Y.; Hong, H. (2004); Robust adaptive control of a class of nonlinear systems with unknown dead-zone, *Autom*, ISSN 0005-1098, 40(3):407-413.
- [13] Yeşildirek, A.; Lewis, F.L. (1995); Feedback linearization using neural networks, *Autom*, ISSN 0005-1098, 31(11):1659-1664.

Performance Analysis of Epidemic Routing in DTN with Overlapping Communities and Selfish Nodes

Y. Wu, S. Deng, H. Huang, Y. Deng

Yahui Wu, Su Deng, Hongbin Huang

Science and Technology on Information Systems Engineering Laboratory
National University of Defense Technology
Changsha, 410073, China
wuyahui@nudt.edu.cn, yahui_wu@163.com, dzyxxx@163.com

Yiqi Deng

University College London
Department of Computer Science
zndxxb@sina.com

Abstract: Routing algorithms in delay tolerant networks (DTN) adopt the store-carry-forward way, and this needs the nodes to work in a cooperative way. However, nodes may not be willing to help others in many applications and this behavior can be seen as *individual selfish*. On the other hand, nodes often can be divided into different communities, and nodes in the same community often have some social ties. Due to these social ties, nodes are more willing to help the one in the same community. This behavior can be seen as *social selfish*. Note that some nodes may belong to more than one community in the real world, and this phenomenon makes the network have overlapping communities. This paper proposed a theoretical model to describe the performance of epidemic routing (ER) in such network. Simulation results show the accuracy of our model. Numerical results show that the selfish nature can make the performance of the routing policy be worse, but those nodes belonging to multi-communities can decrease the impact of the selfish nature in certain degree.

Keywords: Delay Tolerant Networks (DTN), selfish nodes, overlapping communities, epidemic routing, performance analysis.

1 Introduction

At present, there has been a growing interest to study the communication policy for challenged networking applications, such as deep-space exploration [1], vehicular networks [2], mobile social networks [3], etc. In these new environments, the end-to-end connectivity cannot be assumed because the network is quite sparse or nodes are moving fast. That is, a complete path from source to destination does not exist or such a path is highly unstable and may change or break soon after it has been discovered. These networks belong to the general category of Delay Tolerant Networks (DTN) [4]. In traditional Mobile Ad Hoc Networks (MANET), nodes communicate with each other based on the assumption that there exists at least one fully connected path between communication nodes. Therefore, routing policies in MANET cannot be used directly in DTN. In order to overcome the network partitions, nodes of DTN communicate through a store-carry-forward mode. Due to the node mobility, different links come up and down. If the sequence of connectivity graphs over a time interval is overlapped, then an end-to-end path might exist, so the message should be forwarded over the existing link, stored and carried at the next hop until the next link comes up [5].

Many routing policies have been proposed in DTN. According to the number of replicas, these policies can be divided into two classes: that is, single-copy and multi-copy. In the first class, nodes keep only one copy of the message and attempt to forward that copy towards the node which has higher probability to meet the destination, such as the works in [6], [7], etc. Therefore,

how to select the proper relay nodes to carry the copy is critical. In the multi-copy methods, one message may have many replicas and they are transited at the same time to increase the successful ratio [8]- [9]. The core is to select proper relay nodes to store or forward these copies. Therefore, routing policies in both classes depend on the help of other nodes. However, nodes may not be willing to help others due to the constraint of buffer space or power resources [10]. This behavior can be seen as *individual selfish* [11]. Hui et al. studied the its impact in mobile social network, and they found that mobile social network is robust to the *individual selfish* due to the multiple paths [11]. Then, it was studied further in [12]. There are also some incentive methods to make nodes be cooperative[13]-[14]. On the other hand, nodes can be divided into different communities according to their interesting, citation relation, etc. Obviously, nodes in the same community often have some social ties, and they are more willing to help each other. This behavior can be seen as *social selfish*. Li et al. proposed this behavior for the first time [15]. Its impact on the routing performance of ER was explored in [16], and then they studied the impact of both *individual selfish* and *social selfish* on multicasting application [17]. However, above works failed to consider the fact that some nodes may belong to more than one community which is common in the real world [18].

In this paper, we studied the routing performance of ER in DTN with overlapping communities and selfish nodes by the Markov process for the first time. At present, many researchers are interesting in ER algorithm [19]. For example, the performance of ER based on the sparsely exponential graph was studied in [20], and the problem was explored again with heterogeneous nodes [21]. The performance of two-hop relay routing (a special case of ER) under limited packet lifetime was studied in [22]. The authors in work [23] studied the routing performance with contention. In addition, some works begin to study how to decrease the energy consumption of ER. Authors in [24] proposed the optimal probabilistic forwarding policy under a fluid model approximation, and they proved that the optimal policy is the threshold form. Then, they addressed the problem of online estimation of optimal policies in [25], and explored the problem with heterogeneous nodes in [26]. The optimal forwarding problem with multiple destinations was proposed in [27]. Li et al. designed an optimal relaying scheme for DTN, which considers nodes' heterogeneous contact rates and delivery costs when selecting relays to minimize the delivery cost while satisfying the required message delivery probability [28]. The optimal control problem with dead nodes was proposed in [29]. However, to our best knowledge, none of the works considered the problem as ours.

2 Network Model

The set of nodes in the network is denoted by \mathbf{V} . Besides the source S and destination D , every node belongs to at least one of the two communities, which are denoted by $C1$ and $C2$, respectively. It is easy to see that the source and destination may belong to any class. For simplicity, we assume that D belongs to $C1$ and S belongs to $C2$. In fact, our work can be extended to other cases easily. Nodes other than the destination can be seen as relay nodes. The number of relay nodes in the first class is M and the second class has N nodes. Due to the overlap of the communities, there are $O \leq \min \{M, N-1\}$ nodes belonging to both classes. Therefore, there are totally $V=M+N+1-O$ nodes.

The link exists between two nodes only when they come into the transmission range of each other, which means a contact, so the mobility of the users is critical. In this paper, we assume that the occurrence of contacts between two nodes follows a Poisson distribution, which is found in many well-known mobility models, such as random waypoint and random direction [30]. This assumption also has been checked by certain real motion traces [31]. Therefore, we can assume that the inter-meeting time between two contacts follows an exponential distribution

with parameter λ .

Nodes may not be willing to help others due to the individual selfish nature. In this paper, we assume that nodes in $C1$ help the one in the same class with probability p_1 , and nodes in $C2$ help the one in the same class with probability p_2 . On the other hand, nodes is social selfish, so we assume that nodes in $C1$ help nodes belonging to $C2$ with probability p_{12} , and nodes in $C2$ help nodes belonging to $C1$ with probability p_{21} . In fact, many papers used this mode to denote the selfish nature of nodes [15], [16] [17], etc. Because nodes are more willing to help the one in the same community, we have $p_1 > p_{12}$ and $p_2 > p_{21}$. In addition, for any two nodes i and j which belong to $C1$ and $C2$ at the same time, we assume that they communicate with each other with probability $p \geq \max\{p_1, p_2\}$. This assumption is based on the observation that nodes having more common hobbies often have much closer relationship. Therefore, they are more willing to help others. For simplicity, we assume $p = \max\{p_1, p_2\}$ in this paper. That is, if $p_1 > p_2$, nodes i and j communicate with probability p_1 , or with probability p_2 .

3 Data Dissemination Process and Performance Analysis

Now, we begin to explore the data dissemination process based on the ER algorithm. First, we give a new classification of the nodes. In particular, nodes only belonging to $C1$ are denoted by $C11$, and nodes just belonging to $C2$ are denoted by $C22$. Nodes belonging to both $C1$ and $C2$ are denoted by $C12$. Therefore, class $C11$ has $M-O$ relay nodes, class $C22$ has $N-O$ relay nodes, and class $C12$ has O relay nodes. A snapshot of the network can be seen in Figure 1.

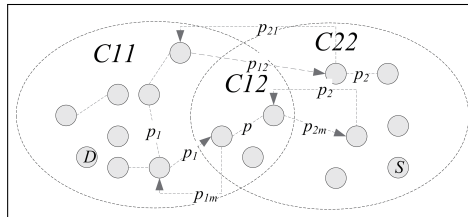


Figure 1: A snapshot of the network

The dashed lines in Figure 1 mean that the link between the nodes is opportunistic. The caption on the line denotes the forwarding probability. Specially, it denotes the forwarding probability from the starting point to the end point. On the other hand, we have $p_{1m} = \max\{p_1, p_{21}\}$ and $p_{2m} = \max\{p_2, p_{12}\}$. That is, for any node i in $C11$ and j in $C12$, because j takes i as a friend, it forwards toward i with probability p_1 . However, node j also belongs to $C2$, if nodes in $C2$ are altruism, node j may forward toward i with probability p_{21} which is bigger then p_1 . Therefore, j should forward to i with probability p_{1m} . We can get the meaning of p_{2m} according to above analysis easily.

3.1 Date Dissemination Model

Let $X(t)$ denote the number of relay nodes in class $C11$ which is carrying data at time t (not including D), $Y(t)$ denote the number of nodes carrying data in $C22$ (including S), and $Z(t)$ denote the corresponding number in $C12$. Therefore, the state of the network at time t can be denoted as $(X(t), Y(t), Z(t))$, and there are totally $(M-O+1)(N-O+1)(O+1)$ transient states. When the destination gets data, the transmission stops, and this state can be seen as the absorption state which is denoted by Dst .

From state $(X(t), Y(t), Z(t))$, the network may change to one of the following four states through one-step transition, that is, $S1 = (X(t)+1, Y(t), Z(t))$, $S2 = (X(t), Y(t)+1, Z(t))$, $S3 = (X(t),$

$Y(t), Z(t)+1$) and Dst . State $S1$ means that one node in $C11$ received data. Obviously, one premise condition of this transition is $X(t) < M-O$, which means that at least one relay node in $C11$ does not receive data at time t . The node in $C11$ which just received data can get the data from nodes in any class. For simplicity, if the node received data from node j , we say that the transition is triggered by j . Obviously, node j may be any node in the network which is carrying data. If j is in class $C11$, one node in $C11$ without data must encounter with node j , and node j is also willing to forward data to it. Because there are $X(t)$ nodes in the class $C11$ carrying data at time t , so there are $M-O-X(t)$ nodes without data in $C11$. Obviously, node j may be any node in the $X(t)$ nodes, and the new node which just received data may be any one in the $M-O-X(t)$ nodes. In addition, nodes encounter with each other according to the exponential distribution with parameter λ , combining the selfish behavior, we know that the transition rate is $\lambda X(t)(M-O-X(t))p_1$. If the transition is triggered by nodes in $C22$, nodes without data in $C11$ must encounter with one node which has received data before in $C22$. Because there are $Y(t)$ nodes in the class $C22$ which is carrying data at time t , and nodes in $C22$ forward to nodes in $C11$ with probability p_{21} , we can know that the transition rate is $\lambda Y(t)(M-O-X(t))p_{21}$. By the same method, we know that if the data comes from $C12$, the transition rate is $\lambda Z(t)(M-O-X(t))p_{1m}$. Now, we can get the total transition rate from state $(X(t), Y(t), Z(t))$ to $S1$ through one-step transition which is shown as follows,

$$(X(t), Y(t), Z(t)) \rightarrow S1, \text{rate } \lambda(M - O - X(t))(X(t)p_1 + Y(t)p_{21} + Z(t)p_{1m}) \quad (1)$$

Similarly, we can get the transition rate from state $(X(t), Y(t), Z(t))$ to $S2$ and $S3$ through one-step transition, which is shown as follows,

$$\begin{aligned} (X(t), Y(t), Z(t)) &\rightarrow S2, \text{rate } \lambda(N - O - Y(t))(X(t)p_{12} + Y(t)p_2 + Z(t)p_{2m}), \\ (X(t), Y(t), Z(t)) &\rightarrow S3, \text{rate } \lambda(O - Z(t))(X(t)p_1 + Y(t)p_2 + Z(t)p) \end{aligned} \quad (2)$$

If the network comes into Dst from state $(X(t), Y(t), Z(t))$, the destination D must receive data. According to above analysis and the forwarding probability in Figure 1, we can get,

$$(X(t), Y(t), Z(t)) \rightarrow Dst, \text{rate } \lambda(X(t)p_1 + Y(t)p_{21} + Z(t)p_{1m}) \quad (3)$$

Let \mathbf{Q} denote the generate matrix which is defined as follows,

$$\mathbf{Q} = \begin{pmatrix} \mathbf{T} & \mathbf{R} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad (4)$$

The elements in the matrix are different. \mathbf{T} is a sub-matrix and it denotes the rate of the transition from one transient state to another. So the number of the rows and columns of the matrix is both $(M-O+1)(N-O+1)(O+1)$. \mathbf{R} is a column vector with $(M-O+1)(N-O+1)(O+1)$ elements and it denotes the rate of the transition from one transient state to the absorbing state Dst . The left $\mathbf{0}$ is a row vector with $(M-O+1)(N-O+1)(O+1)$ elements and it denotes the rate of the transition from Dst to any transient state. The right $\mathbf{0}$ is a vector with only one element and it denotes the rate of the transition from Dst to Dst . According to Equations (1), (2) and (3), we can get every element of \mathbf{Q} . For example, from state (x, y, z) , we have,

$$\begin{cases} T(x+1, y, z|x, y, z) = \lambda(M - O - x)(xp_1 + yp_{21} + zp_{1m}), \\ T(x, y+1, z|x, y, z) = \lambda(N - O - y)(xp_{12} + yp_2 + zp_{2m}), \\ T(x, y, z+1|x, y, z) = \lambda(O - z)(xp_1 + yp_2 + zp), \\ R(Dst|x, y, z) = \lambda(xp_1 + yp_{21} + zp_{1m}), \\ R(others|x, y, z) = 0 \end{cases} \quad (5)$$

Symbol *others* may be any state other than $(x+1, y, z)$, $(x, y+1, z)$, $(x, y, z+1)$ and Dst .

3.2 Performance Analysis

First, we define the one-step transition probability matrix \mathbf{P} which can be got from the generator matrix \mathbf{Q} easily. For example, the transition probability from state i to j is $P(j|i)$, which is an element of \mathbf{P} . Each row of \mathbf{Q} represents the transition rate from one state to others. Therefore, the sum of all elements in one row denotes the rate of leaving the current state. For example, given state SS , the rate of leaving this state denoted by $speed(SS)$ can be shown as,

$$speed(SS) = \sum_{i \in Sspace} Q(i|SS) \quad (6)$$

Symbol $Sspace$ represents the set of all valid states and $Q(i|SS)$ is one element in \mathbf{Q} which represents the transition rate from SS to i . Now, we can get the probability of the transition.

$$P(i|SS) = Q(i|SS)/speed(SS), i \in Sspace \quad (7)$$

Let $DT(k)$ denote the average delivery delay till D received data, starting from state $k=(x, y, z)$. Obviously, we have $DT(Dst)=0$. Similarly, let $ST(k)$ denote the residence time in state k and we also have $ST(Dst)=0$. For any transient state k , we have $speed(k)>0$ and $ST(k)=1/speed(k)$. By conditioning on the one-hop transition out of the current state, we have

$$\begin{aligned} DT(k) &= \sum_{j \in Sspace - \{k\}} P(j|k)DT(j) + ST(k) \\ &= \sum_{j \in Sspace} P(j|k)DT(j) + ST(k) - P(k|k)DT(k) \\ &= \sum_{j \in Sspace} P(j|k)DT(j) + ST(k) \end{aligned} \quad (8)$$

Define \mathbf{DT} as a column vector of the average delivery delay starting from any valid transient state, and \mathbf{ST} also a column vector of the residence time. Then, we can obtain,

$$\mathbf{DT} = \mathbf{P} * \mathbf{DT} + \mathbf{ST} \Rightarrow \mathbf{DT} = (\mathbf{I} - \mathbf{P})^{-1} \mathbf{ST} \quad (9)$$

Because only the source has data at the beginning, we know that the initial state is $initialstate=(0, 1, 0)$. Therefore, the average delivery delay is $\mathbf{DT}(initialstate)$.

Now, we begin to compute the average energy cost using similar method. Here, we only consider the energy cost in the transition process. As descried in [24], [25] and [26], the energy cost is proportional to the number of transmissions which contain both the forwarding and receiving process. In this paper, we use the number of transmissions to denote the energy cost simply. Let $ER(k)$ denote the average energy cost till D received data, starting from state $k=(x, y, z)$, obviously $ER(Dst)=0$. According to above analysis, from state k , the network may come into any state of the following four states: $k1=(x+1, y, z)$, $k2=(x, y+1, z)$, $k3=(x, y, z+1)$ and Dst . Obviously, if the network changes into one of them, one node must forward data and the other one must receive data. That is, if the network changes state, there is one transmission. Therefore, we can obtain,

$$\begin{aligned} ER(k) &= P(k1|k)(1 + ER(k1)) + P(k2|k)(1 + ER(k2)) \\ &\quad + P(k3|k)(1 + ER(k3)) + P(Dst|k)(1 + ER(Dst)) \\ &= \sum_{j \in Sspace} P(j|k)(1 + ER(j)) = \sum_{j \in Sspace} P(j|k) + \sum_{j \in Sspace} P(j|k)ER(j) \quad (10) \\ &= 1 + \sum_{j \in Sspace} P(j|k)ER(j) \end{aligned}$$

Define \mathbf{ER} as the corresponding column vector. Equation (10) can be changed to the following equation.

$$\mathbf{ER} = \mathbf{P} * \mathbf{ER} + \mathbf{e} \Rightarrow \mathbf{ER} = (\mathbf{I} - \mathbf{P})^{-1} \mathbf{e} \quad (11)$$

Symbol \mathbf{e} is a column vector and every element in it equals to 1. Therefore, the average energy cost starting from state $initialstate=(0, 1, 0)$ is $\mathbf{ER}(initialstate)$.

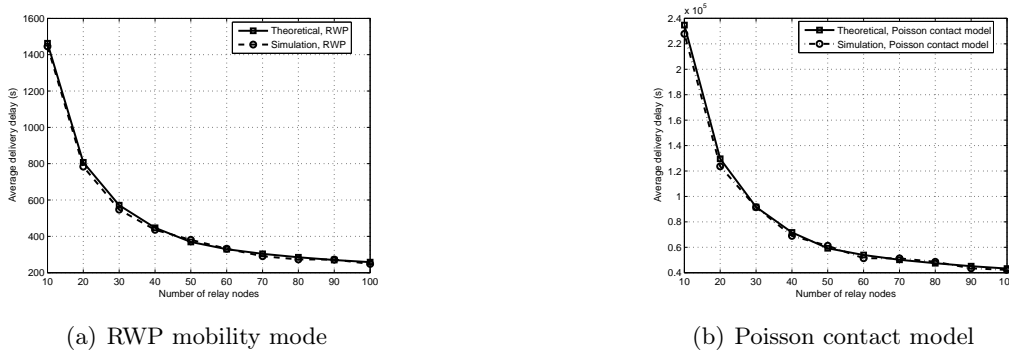


Figure 2: Theoretical and simulation result comparison of the average delivery delay

4 Simulation and Numerical Results

4.1 Simulation Results

In this section, we will check the accuracy of our theoretical model, and we run several simulations using the Opportunistic Network Environment (ONE) simulator [32]. The simulation is based on both synthetic mobility model and real-world-based scenarios. The synthetic model is the famous Random Waypoint (RWP) mobility model. In this model, the simulation terrain is $1000m \times 1000m$, and the speed varies from 0.5 to 1.25m/s. The transmission range is 2m. For the real-world-based scenario, we use the Poisson contact model. Specially, we have $\lambda = 3.71 \times 10^{-6} s^{-1}$. As shown in [17] and [33], this value is obtained from the vehicle model, which is based on real motion traces from about 2100 operational taxis for about one month in Shanghai city collected by GPS. Authors of [34] proposed a least-fitting method to identify the exponential parameter and find that above value is well proper. For the theoretical model related parameters, we set $p_1=0.5$, $p_{12}=0.2$, $p_2=0.8$ and $p_{21}=0.1$. As described above, there are $totalnum=M+N-O$ relay nodes in the network. Without loss of generality, we set $M=N$ and $O=0.2totalnum$. Through let the number of relay nodes increase from 10 to 100, we get the results in Figure 2.

From the result we can see that the average deviation between the theoretical results and the simulation is very small. For example, the deviation is about 2.8% for the RWP mobility model and 4.6% for the Poisson contact model. This demonstrates the accuracy of our theoretical model. Then, we will use the theoretical results obtained by our model to evaluate the performance in different cases.

4.2 Performance Analysis with Numerical Results

First, we will explore the impact of the overlap between communities. Here, we increase the value of O continuously till reaching to $totalnum$. Let the ratio $O/totalnum$ increase from 0.1 to 1. Other settings are the same as that in RWP model in the simulation. The numerical results are shown in Figure 3 when the number of relay nodes equals to 20, 40 and 80, respectively.

Figure 3 shows that if there are more nodes belonging to both communities at the same time, the average delivery delay will be smaller. For example, when $N=20$, the average delivery delay is reduced by about 79.2% when the value of $O/totalnum$ increases from 0 to 1. However, the value of $O/totalnum$ has little influence on the average energy cost.

Then, we will explore the impact of the overlap when the selfish level is different. First, we explore the case with different *social selfish* level, so we can assume that nodes in the same community are altruism for each other, that is $p_1=p_2=1$. The total number of relay nodes

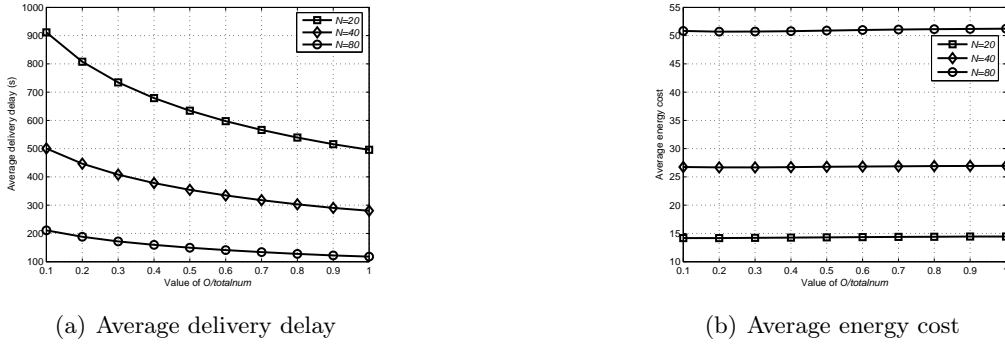


Figure 3: Impact of the overlap phenomena with different number of relay nodes

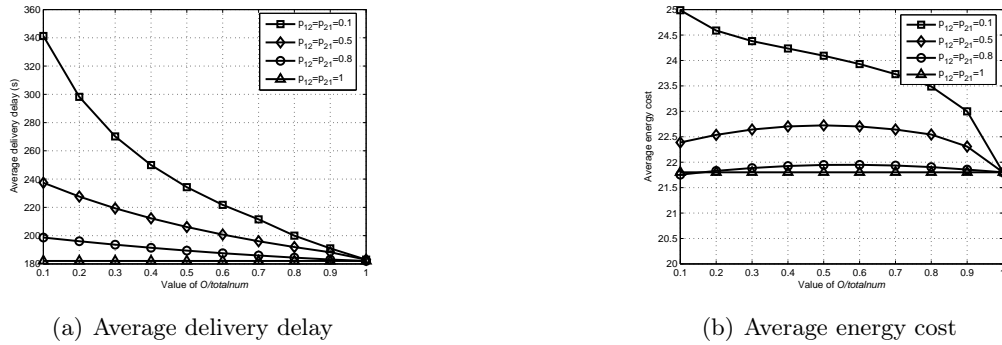


Figure 4: Impact of the overlap phenomena with different level of social selfish

$totalnum$ is 40, and we have $M=N$. Other settings are also the same as that in RWP model. We set $p_{12}=p_{21}$, and let their value equal to 0.1, 0.5, 0.8 and 1, respectively. Through letting $O/totalnum$ increase from 0.1 to 1, we can get Figure 4.

From Figure 4 we can see that with fixed *social selfish* level, the average delivery delay is decreasing with the increasing of the nodes in $C12$ when $p_{12}=p_{21}<1$. When $p_{12}=p_{21}=1$, every node is altruism, so the overlap between communities cannot have any impact. This result also shows that the smaller of p_{12} ($p_{12}=p_{21}$), the bigger of the decreasing ratio will be. In addition, the difference of the delivery delay with different *social selfish* level is decreasing with $O/totalnum$, and they have the same value when $O/totalnum=1$. Figure 4(b) is a surprising result, and it shows that the average energy cost is not monotonous with $O/totalnum$. This demonstrates the complex correlation between the overlap and the social selfish behavior. When the cooperative level is small, for example when $p_{12}=p_{21}=0.1$, the average energy cost is decreasing with $O/totalnum$. This is because that with the increasing of $O/totalnum$, the average delivery delay decreasing rapidly (see Figure 4(a)), data has less time to spread further. However, when nodes are more cooperative, the average energy cost is first increasing, and then begins to decrease with $O/totalnum$. In this case, though the average delivery delay still decreases with $O/totalnum$, the increasing degree of the data spreading speed is much bigger. Therefore, the energy cost increases in some degree, but the impact of the increasing of the data spreading speed becomes smaller when $O/totalnum$ is big enough. In fact, the fluctuation of the average energy cost is very small under different value of $O/totalnum$. Therefore, if there are more nodes belonging to more community, the network can get better performance without much increasing of the energy cost.

Now, we want to explore the results with different *individual selfish* level when the *social*

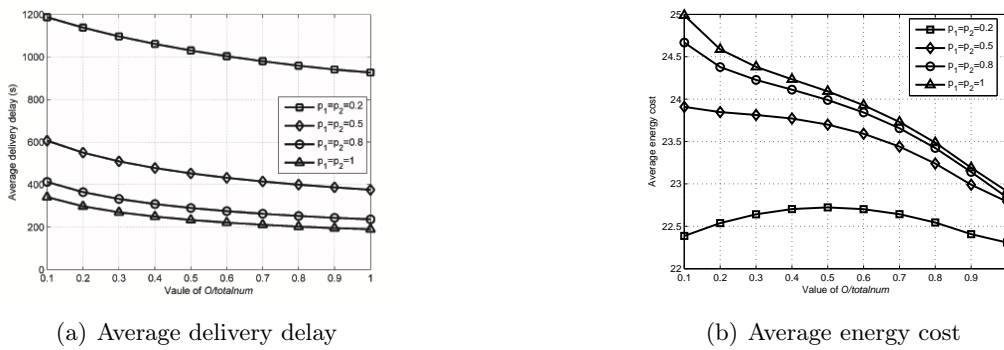


Figure 5: Impact of the overlap phenomena with different level of individual selfish

selfish level is fixed. Here, we assume that nodes in one community helps the one in other community with probability 0.1. Other settings are the same as that in Figure 4. We give the numerical results when $p_{12} = p_{21} = 0.2, 0.5, 0.8$ and 1, respectively. Let $O/totalnum$ increase from 0.1 to 1, we can get Figure 5. The result also shows that the average delivery delay decreases with $O/totalnum$. The average energy cost has similar changing rule as that in Figure 4(b). This further demonstrates that the overlap is good for the network.

5 Conclusions

This paper explored the performance of ER algorithm in DTN which has overlapping communities and selfish nodes, and a theoretical model based on the Markov process was proposed. Simulation results show the accuracy of the model. Numerical results show that the overlap between communities can improve the performance without much increasing of the energy cost.

Bibliography

- [1] S. Haoliang, L. Lixiang and H. Xiaohui, A network coding based DTN convergence layer reliable transport mechanism over interplanetary networks, *International Journal of Computers, Communications and Control*, vol.6, no.2, pp.236-245, 2011.
- [2] A. Rahim, Z. S. Khan, F. B. Muhaya, M. Sher and M. K. Khan, Information sharing in vehicular adhoc network, *International Journal of Computers, Communications and Control*, Vol.5, No.5, pp.892-899, 2010.
- [3] W. Gao, Q. Li, B. Zhao and G. Cao, Multicasting in delay tolerant networks: a social network perspective, *In Proc. ACM MobiHoc*, 2009.
- [4] K. Fall, A delay-tolerant network architecture for challenged internets, *In Proc. ACM SIGCOMM*, 2003.
- [5] T. Spyropoulos, T. Turletti and K. Obrazcka, Routing in delay tolerant networks comprising heterogeneous populations of nodes, *IEEE Transaction on Mobile Computing*, vol. 6, no. 8, 2009.
- [6] T. Spyropoulos, K. Psounis and C. Raghavendra, Efficient routing in intermittently connected mobile networks: the single-copy case, *ACM/IEEE Transaction on Networking*, 2008.

-
- [7] Z. Guo, B. Wang and J. -H. Cui, Prediction assisted single-copy routing in underwater delay tolerant networks, *In Proc. IEEE Globecom*, 2010.
 - [8] E. Bulut, Z. Wang and B. Szymanski, Cost effective multi-period spraying for routing in delay tolerant networks, *ACM/IEEE Transaction on Networking*, vol. 18, no. 5, 2010.
 - [9] W. Gao, G. Cao, On exploiting transient contact patterns for data forwarding in delay tolerant networks, *In Proc. IEEE ICNP*, 2010.
 - [10] G. Resta, P. Santi, The effects of node cooperation level on routing performance in delay tolerant networks, *In Proc. IEEE SECON*, 2009.
 - [11] P. Hui, K. Xu, V. O. K. Li, J. Crowcort, V. Latora and P. Lio, Selfishness, altruism and message spreading in mobile social networks, *In Proc. IEEE NetSciCom*, 2009.
 - [12] K.Xu, P. Hui, V. O. K. Li, J. Crowcort, V. Latora and P. Lio, Impact of altruism on opportunistic communications, *In Proc. IEEE ICUFN*, 2009.
 - [13] R. Lu, X. Lin, H. Zhu, X. Shen and B. Preiss, Pi: a practical incentive protocol for delay tolerant networks, *IEEE Transactions on Wireless Communications*, vol.9, no.4, 2010.
 - [14] T. Ning, Z. Yang, X. Xie and H. Wu, Incentive-aware data dissemination in delay-tolerant mobile networks, *In Proc. IEEE SECON*, 2011.
 - [15] Q. Li, S. Zhu and G. Cao, Routing in socially selfish delay tolerant network, *In Proc. IEEE INFOCOM*, 2010.
 - [16] Y. Li, P. Hui, D. Jin, L. Su and L. Zeng, Evaluating the impact of social selfishness on the epidemic routing in delay tolerant networks, *IEEE Communication Letters*, vol.14, no.11, pp.1026-1028, 2010.
 - [17] Y. Li, G. Su, D. O. Wu, D. Jin, L. Su and L. Zeng, The impact of node selfishness on multicasting in delay tolerant networks, *IEEE Transactions on Vehicular Technology*, vol.60, no.5, 2011.
 - [18] N. P. Nguyen, T. N. Dinh, S. Tokala, M. T. Thai, Overlapping communities in dynamic networks: their detection and mobile applications, *In Proc. ACM Mobicom*, 2011.
 - [19] A. Vahdat, D. Becker, Epidemic routing for partially-connected ad hoc networks, *Technical Report, Duke University*, 2000.
 - [20] X. Zhang, G. Neglia, J. Kurose and D. Towsley, Performance modeling of epidemic routing, *Computer Networks*, vol. 51, no. 10, pp. 2867-2891, 2007.
 - [21] Y. K. Ip, W. -C. Lau and O. -C Yue, Performance modeling of epidemic routing with heterogeneous node types, *In Proc. IEEE ICC*, pp. 219-224, 2008.
 - [22] A. Al-Hanbali, P. Nain and E. Altman, Performance of ad hoc networks with two-hop relay routing and limited packet lifetime, *In Proc. Valuetools*, 2006.
 - [23] A. Jindal, K. Psounis, Contention-aware performance analysis of mobility-assisted routing, *IEEE Transaction on Mobile Computing*, vol. 8, no. 2, pp. 145-161, 2009.
 - [24] E. Altman, T. Basar and F. D. Pellegrini, Optimal monotone forwarding policies in delay tolerant mobile ad-hoc networks, *In Proc. ACM Inter-Perf*, 2008.

- [25] E. Altman, G. Neglia, F. D. Pellegrini and D. Miorandi, Decentralized stochastic control of delay tolerant networks, *In Proc. IEEE INFOCOM*,2009.
- [26] F. D. Pellegrini, E. Altman and T. Basar, Optimal monotone forwarding policies in delay tolerant mobile ad hoc networks with multiple classes of nodes, *In Proc. WiOpt*,2010.
- [27] C. Singh, A. Kumar, R. Sundaresan and E. Altman, Optimal forwarding in delay tolerant networks with multiple destinations, *In Proc. WiOpt*,2011.
- [28] Y. Li, Z. Wang, D. Jin, L. Su, L. Zeng and S. Chen, Optimal relaying in heterogeneous delay tolerant networks, *In Proc. IEEE ICC*,2011.
- [29] M. Khouzani, S. Sarkar and E. Altman, Optimal control of epidemic evolution, *In Proc. IEEE INFOCOM*,2011.
- [30] R. Groenevelt, P. Nain and G. Koole, The message delay in mobile ad hoc networks, *Performance Evaluation*,2005.
- [31] T. Karagiannis, J. -Y. L. Boudec and M. Zojnovic, Power law and exponential decay of inter contact times between mobile devices, *In Proc. ACM Mobicom*,2007.
- [32] A. Keranen, J. Ott, and T. Karkkainen, The ONE simulator for DTN protocol evaluation, *In Proc. SIMUTOOLS*,2009.
- [33] S. J. U. Traffic information grid team, Grid Computing Center, Shanghai Taxi Trace Data. <http://wirelesslab.sjtu.edu.cn/>
- [34] H. Zhu, L. Fu, G. Xue, Y. Zhu, M. Li and L. Ni, Recognizing exponential inter-contact time in VANETs, *In Proc. IEEE INFOCOM*,2010.

A Quick Location Method for High Dynamic GNSS Receiver Based on Time Assistance

P. Wu, S. Jing, W. Liu, F. Wang

Peng Wu*, Shourang Jing,
Wenxiang Liu, Feixue Wang

College of Electronic Science and Engineering,
National University of Defense Technology, Changsha 410073, China
wp4nnc@gmail.com, hanchongjsr@163.com,
liuwenxiang8888@163.com, wangfeixue365@sina.com

*Corresponding author: wp4nnc@gmail.com

Abstract: Traditional A-GPS positioning method when quickly calculate a position, need a condition that the approximate position must not exceed 150km, otherwise the calculation will be very complex. This paper proposes a time-assisted fast positioning method for high dynamic GNSS receiver, effectively solving the problem of large search calculation in traditional method, even if exact position is unknown after the signal is recaptured. According to the known auxiliary time information and implied elevation information, this paper put forwards a custom coordinate system for building two-dimensional search space, which could reduce the number of search-dimensions. It proposes a search method based on receiver clock calculated by analyzing the influence of time auxiliary accuracy. By using GPS ephemeris data provided by the IGS, it builds a simulation environment and analyzes the influence of different preferred satellites based on the custom coordinate system on the calculation, and thus puts forward a principle for choosing the preferred satellites. Simulation examples show that through the rational combination of satellites to create a custom coordinate system, and when time auxiliary accuracy is less than 60us, the calculation can 100% guarantee to restore a complete satellite signal emission time and obtain an accurate position.

Keywords: Global Navigation Satellite System(GNSS), signal transmission time recovery, high dynamic, Time to First Fix(TTFF).

1 Introduction

Since the low speed of the message affects the speed of frame synchronization [1], the traditional non-auxiliary GPS receiver can not obtained the complete time of satellite signal emission instantly, which will affect TTFF (Time to First Fix). Take the GPS L1 as an example, from catching to accomplishing code phase synchronization of several satellites, ordinary receivers take one or two seconds to receive the time of satellite signal emission. But it takes about fifteen seconds to accomplish bit and frame synchronization by message, and receivers can not locate before frame synchronization completed. Through auxiliary methods [2, 3], before frame synchronization, receivers can recover the milliseconds of the signal emission to accelerate TTFF.

However, there are some conditions when using auxiliary GPS, that is only when the error of location estimation is not beyond 150km, can the receiver recover instantly the milliseconds of the satellite and calculate the exact position of the receiver [4-6]. When the receiver is in high dynamics, such that when the design speed reaches 900m/s, and when it lost the signal for a while, such that when the off power and loss of lock time is beyond 167s, the scope of the general location will be beyond 150km, which will increase the work of calculation. Literature [7, 8] proposes to reduce the calculation in searching through narrowing the searching area by visibility of satellite and through removing unreasonable values by calculation. But this method is still complex, for it is based on the calculation of the receiver's dimensional position and clock error,

in which the dimensional search does not make use of information hidden on the surface of the earth. Literature [9–11] puts forward a kind of location by atomic clock assisted.

This paper puts forward a new idea that when calculating the general location, the four unknown numbers (the three dimensions and clock error) can be reduced to two (the two dimensions of the receiver on the earth), which is a combination of a custom coordinate system and the hidden information on the earth of the receiver. When calculating the two dimensions, there is new two-dimensional searching space and a searching method based on receiver clock calculated.

2 Building the Searching Space

With auxiliary time information and elevation information, search of the user’s location is a search of the two dimensions in the horizontal level. To be convenient, the two dimensions in the horizontal level is redefined by the custom polar coordinate.

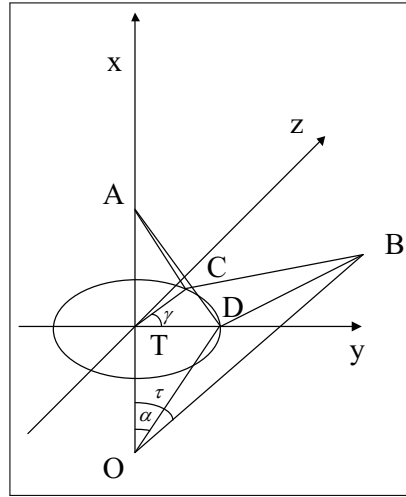


Figure 1: Custom coordinate system

As the graph shows, O represents the earth, with A and B two satellites. Satellite A and satellite B who build the custom coordinate system are called the preferred satellites for the convenience, while other satellites are called validate satellites.

In the custom right-angle coordinate system, O is origin, \vec{OA} is axis Z, axis Y is vertical toward plat \vec{OAB} , and axis X is vertical toward plat \vec{OYZ} and on the left hand of YZ. The unit vector $\vec{e}_x, \vec{e}_y, \vec{e}_z$ of X, Y, Z can be stated as $\vec{e}_x = \frac{\vec{e}_y \times \vec{e}_z}{|\vec{e}_y \times \vec{e}_z|}, \vec{e}_y = \frac{\vec{OA} \times \vec{OB}}{|\vec{OA} \times \vec{OB}|}, \vec{e}_z = \frac{\vec{OA}}{|\vec{OA}|}$. “ \times ” represents outer calculation of the vector. Therefore, the arbitrary point C can be represented as $(\vec{OC} \cdot \vec{e}_x, \vec{OC} \cdot \vec{e}_y, \vec{OC} \cdot \vec{e}_z)$ in the system.

Suppose satellite A arrives at the receiver signal \vec{AC} , whose intersection with plat XY is D. Then the physical meaning of the circle \vec{CTD} they form is a collection of the same transmission time delay of the satellite signal when it arrives at the earth. \vec{BD} represents the shortest distance when satellite B arrives at the receiver, on the condition that time delay of satellite A is in certainty. Then the problem of satellite B’s obscure time delay is to calculate the length of \vec{BD} . R is the distance between the satellite and the earth center, such as \vec{OA} . τ is the intersection angle of two satellites against the earth center. r is the distance between the receiver and the earth center, such as \vec{OD}, \vec{OC} . The result is a combination of the radius of the earth semi-major and the elevation the receiver estimates.

We can see that when r is known by the estimated elevation, the receiver in the horizontal level is decided by angle α and angle γ . It will be clearer when we cite the polar coordinate in three-dimension. The referential value is the length between the point and O. The first referential angle is the intersection angle α of \overline{AOD} , and the second referential angle is the intersection angle γ of \overline{CTD} against the surface of circular cross-section. Then the coordinates of A, B, C in the space can be presented as $A : (R, 0, 0)$, $B : (R \cos \tau, R \sin \tau, 0)$, $C : (r \cos \alpha, r \sin \alpha \cos \gamma, r \sin \alpha \sin \gamma)$. Then:

$$\overline{AC} = \sqrt{R^2 + r^2 - 2rR \cos \alpha}, \quad \overline{BC} = \sqrt{R^2 + r^2 - 2rR \cos \alpha \cos \tau - 2rR \sin \alpha \sin \tau \cos \gamma}$$

3 Search Method on the Basis of Clock Error

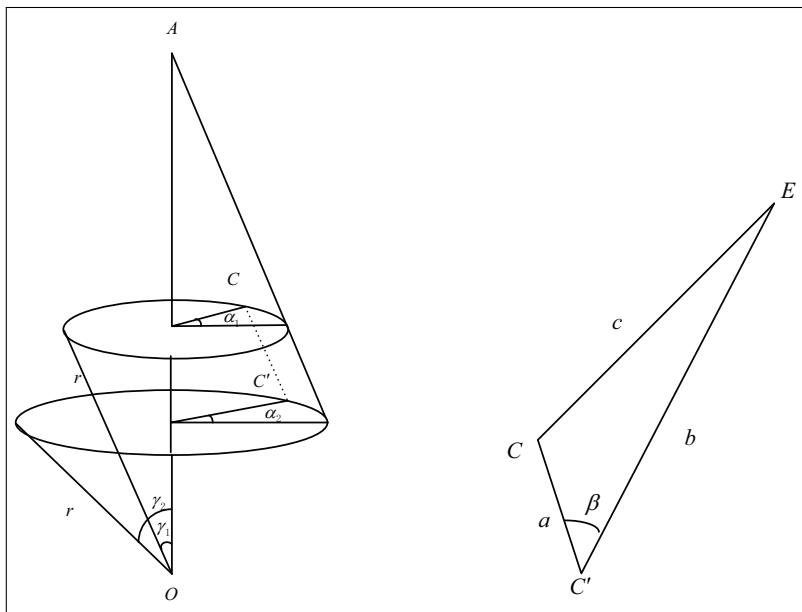


Figure 2: Influence on receiver position by time bias

As the graph shows, C is the real location of the receiver, with coordinate (r, α_1, γ_1) . C' is the deviated location caused by clock error, with coordinate (r, α_2, γ_2) . According to the custom coordinate system, coordinate values of C and C' in the space are:

$$C : (r \cos \alpha_2, r \sin \alpha_2 \cos \gamma_2, r \sin \alpha_2 \sin \gamma_2), \quad C' : (r \cos \alpha_1, r \sin \alpha_1 \cos \gamma_1, r \sin \alpha_1 \sin \gamma_1).$$

Suppose $\Delta\alpha = \alpha_2 - \alpha_1$, $\Delta\gamma = \gamma_2 - \gamma_1$, local time error is Δb , error of the direction vector $\overrightarrow{C'C}$ is

$$\begin{cases} \vec{x} = \frac{\cos \gamma_1 |AC|}{R \tan \alpha_1} \Delta b - \frac{|AC|}{R \sin \tau} \left(1 - \frac{|AC|}{|BC|} \cos \tau + \frac{|AC|}{|BC|} \cot \alpha_1 \sin \tau \cos \gamma_1 \right) \Delta b \\ \vec{y} = \frac{\sin \gamma_1 |AC|}{R \tan \alpha_1} \Delta b + \frac{|AC|}{R \sin \tau \tan \gamma_1} \left(1 - \frac{|AC|}{|BC|} \cos \tau + \frac{|AC|}{|BC|} \cot \alpha_1 \sin \tau \cos \gamma_1 \right) \Delta b \\ \vec{z} = -\frac{|AC|}{R} \Delta b \end{cases}$$

Namely, $\overrightarrow{C'C}$ is the only value that corresponds Δb , so it can be presented as $\overrightarrow{C'C} = f(\Delta b)$.

From the validate satellite E, we can search for packs in the space according to the geometric relations and determine the vectors $\overrightarrow{CC'}$ and $\overrightarrow{EC'}$ as well as the included angle β , thus working out \overrightarrow{EC} . Assuming that these vectors are a, b, c , then we have the following expression:

$$\begin{cases} a = |CC'| = f(\Delta b) \\ b = |EC'| = c + \Delta b \\ c = \sqrt{a^2 + b^2 - 2ab \cos \beta} \end{cases}$$

The only unknown in the equation set Δb can be solved by simultaneous equation. After working out the Δb , we can compute the variance of clock error packs and make the minimum variance as the truth value due to the uniformity of Δb of different validate satellites.

4 Experiment and Analysis

For the experiment, GPS constellation of IGS ephemeris is selected, the time being 25 December 2011. When the receiver is at a position of N28.2°E112.9°, a total of ten satellites are visible and their distribution and position are as follows:

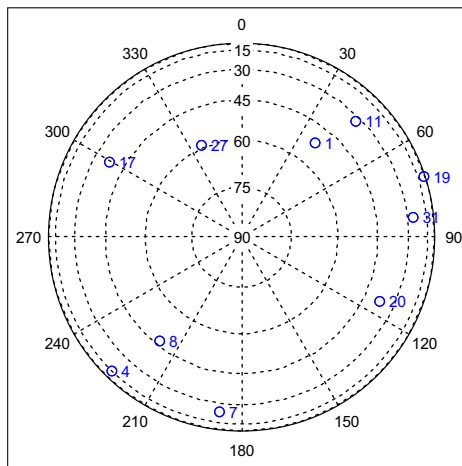


Figure 3: Sky charts of visual satellites

The chosen sphere of movement of the receiver is 10°N~30°N, 120°E~140°E. Simulation test imitates the movement of the receiver within this sphere at a pace of 1° per movement. Height precision error is set at no more than 10km, time auxiliary error at 0~200us. In order to illustrate the impact on the calculation by the two satellites, we select two pacts: the first pact is of No. 1 and No. 4 satellites, and the second pact No. 8 and No. 20.

In order to illustrate the impact of different pacts of satellites on algorithms, we compare two pacts to show the success rate distribution of both ways of calculation when there are 2, 4, 6 and 8 satellites involved. The results are as follows:

The above result shows that:

1. Using auxiliary time information, the two pacts can both lead to the launching time of the satellite correctly. But the success rate, as it is subject to the precision of the time auxiliary, is more or less reduced as the error increases.
2. Normally, the more the validate satellites the higher success rate. In reality, however, the success rate of pacts of two satellites is remarkably lower than those of pacts of more than

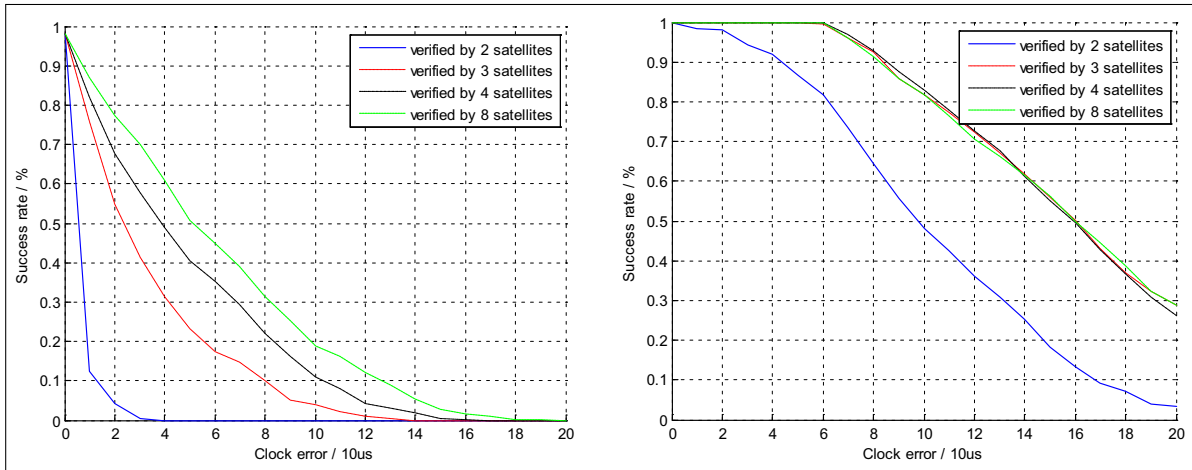


Figure 4: Success rate of algorithm in different conditions

two satellites. It means that when the number of validate satellites is more than two, an increase in the number of validate satellites does not lead to an obvious increase in success rate.

3. The algorithm of the first pact of first choice satellites does not guarantee a success rate of 100%. Even if the clock error is 0, the success rate is only as high as 98%. As the clock error increases, the success rate drops quickly enough to render the result virtually useless.
4. The second pact of first choice satellites can yield a success rate of 100% when the clock error is below 60us. When the error increases, the success rate drops at a lower rate than with the first pact. Therefore, it can be applied to cases of a wider range of time errors.

5 Conclusion

To sum up, to safeguard the success rate of algorithm, we can determine the principle of first choice satellite selection. In other words, when selecting preferred satellites, we must work out the area where the receiver can capture the signal of all the satellites according to the distribution of visible satellites. The selection principle is:

1. Select pacts on the same side on the verge of the area. Make sure that the line connecting the first choice satellites is nowhere near the top of the receiver.
2. Under the above principle, select pacts of satellites with greater distances.

The first principle can overcome the situation when PDOP value is infinite. The Second principle is conducive to further reducing PDOP value and increasing the application scope of local time auxiliary precision. It is also good for reducing scope of searching space and reducing calculation load.

Bibliography

- [1] Kaplan, E. D., Hegarty, C. J.; *Understanding GPS: principles and applications*, 2nd ed. Norwood, MA, Artech House, 2006.
- [2] Agarwal, N., Basch, J., Beckmann, P.; Algorithms for GPS operation indoors and downtown, *GPS Solutions*, 6:149-160, 2002.
- [3] Diggelen, F. V.; Indoor GPS theory & implementation, *IEEE Position, Location & Navigation Symposium*, 40-247, 2002.
- [4] HUANG Yiping; GPS receiver with extended hot start capability: US, 7466265, 2008.
- [5] Akopian, D., Syrjärinne, J.; A fast positioning method without navigation data decoding for assisted GPS receivers, *Vehicular Technology*, 58:4640-4645, 2009.
- [6] LI, Jizhong, WU, Muqing; A Positioning Algorithm of AGPS, *International conference on signal processing systems*, 385-388, 2009.
- [7] Sirola N.Syrjärinne, J.; GPS position can be computed without the navigation data, *ION GPS*, 2741-2744, 2002.
- [8] CAO, Hui, YUAN, Hong; Method for Time-of-transmission Recovery Based on Assisted-GPS Positioning, *Chinese J.Space Sci*, 32(3):585-591, 2012.
- [9] Yang Chuan, Wang Yongsheng, SHI Lijian; Study of atom clock aiding GPS positioning, *GNSS World of China*, 4:5-8, 2005.
- [10] John Kitching, Svenja Knappe, Li-Anne Liew, et al. Chip-Scale Atomic Frequency References. *ION GNSS 18th International Technical Meeting of the Satellite Division*, September 2005, Long Beach, CA, 1662-1669.
- [11] SA.45s Chip Scale Atomic Clock Data Sheet. Symmetricom 2013. <http://www.symmetricom.com/products/>.
- [12] WU Peng, XU Bo, LIU Wenxiang, WANG Feixue. GNSS emergency positioning method and research on the accuracy estimation[C].//The forth sector of the Chinese Satellite Navigation Conference, (CSNC). Wuhan, 2013.
- [13] WU Peng, WANG Feixue. The method and analysis of DOP calculation in three-satellite positioning in BD system, *The second sector of the Chinese Satellite Navigation Conference*, (CSNC). Shanghai, 2011.

Scalable Architecture for CPS: A Case Study of Small Autonomous Helicopter

J. Yao, J. An, F. Hu

Jianguo Yao*, Jie An, Fei Hu

School of Software

Shanghai Jiao Tong University, Shanghai, China

800 Dongchuan Road, Minhang, Shanghai 200240, China

*Corresponding author jianguo.yao@sjtu.edu.cn

anjie@sjtu.edu.cn, hufei@sjtu.edu.cn

Abstract: Building a scalable and highly integrated systems is an important research direction and one of key technologies in Cyber-Physical Systems (CPS). Autonomous helicopter is a typical CPS application and its flight presents challenges in flight control system design with scalability. In this paper, we present the integration architecture of hardware and software for the flight control system based TREX 600 helicopter. In order to enhance scalability, the flight control system uses the PC104 and the ARM which is exerted to process the measurement data, including the position, attitude, height etc. The flight control is developed based multi-loop decoupling PI control which is easy to be implemented. Finally, the flight control system is successfully verified in the actual autonomous flight control experiment.

Keywords: Cyber-Physical Systems (CPS); autonomous helicopter; component.

1 Introduction

Cyber-Physical Systems (CPS) consider both cyber parts (e.g., computing and communication) and physical parts (e.g., hardware), as well as their interactions [1]. Autonomous helicopter is one of typical applications in CPS and requires dynamic configuration of system-level performance. Helicopter has the advantage of being able to vertically take-off/landing, hover and lateral flight. Because of its flexible flight maneuver, even in a restricted region helicopters can fly free and hover in the air for a long time. These characteristics make unmanned aerial vehicle (UAV) applicable for many military and civil applications. For example, geographic map detection and aerial photography, security patrols, combat assault, resources exploration, forest fire prevention, film shooting, pesticide spraying etc. UAV has become a popular research topic for many military and university research institutions recently.

The UAV flight control system developed by Stanford University airborne module includes two GPS receiver, one tachometer for wing feedback, two sets of video camera and transmitter; one Pentium-486 processing board, which can realize the automatic hovering and landing [2]. The American Georgia Tech University UAV airborne module includes one DGPS receiver, one heading magnetometer, three axis inertial attitude angular velocity measuring unit, one set of sonar and radar altimeter gets height information, one set of cameras, one PIIPC processing board and wireless card, which not only can manipulate the autonomous flight, can also identify the building and its entrance [3]. American Berkeley University Research of UAV, now can be done independently of the take-off and landing, obstacle avoidance in the complex environment based vision technology, on specific target tracking and anti-tracking, collision detection and multiple machine coordinated flight. Meanwhile, there are researches on UAV flight control system in China, such as Zhejiang University has implemented UAV airborne system which is based DSP-CPLD, the main research of it is in the hardware realization and software interface; Nanjing University of Aeronautics & Astronautics has designed a UAV flight control system which is based 586-Engine micro control module.



Figure 1: Implemented rotorcraft UAV.

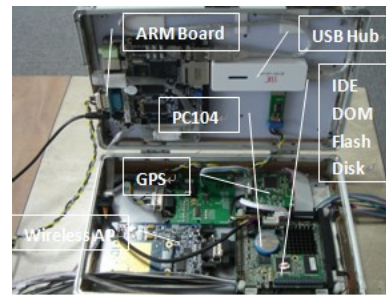


Figure 2: The avionics control system box.

Based on the listed components for a UAV airborne system, we design and assemble a simple prototype UAV helicopter. A TREX 600 radio-controlled helicopter is chosen as the basic flying vehicle. A simple avionic system will be designed. The miniature PC104 [4] [5] computer will be taken as the airborne computer system. ARM will be taken as the sensor data collection controller [6] [7]. The ARM data collection controller not only reduce the burden on the PC104, but also make the airborne system scalable. The full avionic system has been tested successfully on controlling the TREX 600 helicopter model, and the helicopter can realize fixed point hover and autonomous trajectory plan flight.

The outline of this paper is as follows. The components of avionic system are introduced in Section 2 in details. A flight control law which is based multi channel decoupling PID control has been designed and implemented in Section 3. Section 4 show the effect of the overall actual flight control experiment. Finally we summarize some conclusions in Section 5.

2 TREX 600 Helicopter Model Description

We choose a small size model helicopter (i.e., TREX 600) which is available in the market as the basic aircraft of UAV system. TREX 600 is a low cost model helicopter and it can easily to be upgraded. We can buy the individual parts of the model helicopter in market easily.

TREX 600 model helicopter is a high quality toy helicopter in the hobby industry. The control way of main rotor blades is CCPM, three servos are linked to the swashplate to control collective and cyclic angles of attack of the main rotor blades. We upgraded the model to be a UAV helicopter with a complete auto-pilot system, sensors and a necessary communication modem. The manual and autonomous control model is switched by the 7 channel of the receiver. The helicopter itself weights about 1770g, and it can provide an effective load up to 1200g, which is far above our budget on the weight of the avionic system and other components to be upgraded. The fuselage of helicopter is constructed with carbon fiber board, which can well protect the flight control equipment. The protected principle is the helicopter body will break open at the crash moment, and the disintegration of aircraft can reduce the impact of crash. The main rotor blades are replaced with heavy-duty carbon fiber ones, which has the quality of high weight and high hardness. The new rotor blades can accommodate extra payloads. The helicopter is powered by a 600XL brushless electric motor which generate 2.68hp at about 14000rpm. The full length of fuselage is 1200mm, the height of helicopter is 405mm. The main rotor is about 1350mm and the tail rotor is 240mm. The implemented rotorcraft UAV is shown in Fig. 1.

Designing and installing the flight control system is the first step to implement of the UAV helicopter system. It is necessary to take into account the particularity of environment during the design process, strong magnetic field as well as intense vibration interference environment. So we must adopt the shielding and shock-absorption measures. In addition, the weight and

the size of the avionics box are strict limited because of the limited payload capacity. On the other hand, helicopter's center of gravity position has a great influence on the flight. The general center of gravity position is in front of the main spindle 10mm. The standard must be fulfilled both before and after the installation of flight control system, so as not to destroy the balance of the helicopter.

We adopt the design pattern of integrated avionic system [8], which is shown in Fig.2. All the flight control system equipment installed in a $26 \times 16 \times 7$ cm aluminum box, which can play the role of electromagnetic shielding. To try not change the aircraft center of gravity, aluminum box frame mounted on the bottom of aircraft's landing gear. Because of there is no enough room under the original landing gear to install the designed avionics system. While we re-design a landing gear with aluminum alloy and make a larger room under the gear for the control box, the dimensions of it is $30 \times 29 \times 43$ cm. Soft connection is adopted between the new and old landing gear, which can protect the flight control equipment when the helicopter land on the ground. To avoid the disciplinary vibration about 30HZ caused by characteristic of the helicopter, silicon rubber are mounted between the aluminum box and the new land gear. Experiments show that the measures we take can effectively isolate vibration source.

3 Scalable Architecture for Helicopter System

The overall structure of the flight control system shown in Fig.3. Using a PC104 as a flight control computer, it is the core of the system. The input data of the PC104 are flight sensor data and control commands, output data of it is servo control signal PWM signal. One ARM board is used as the flight data collection module, it collect data of each sensor and then send them to PC104 flight control system via Ethernet. Ground station PC send PC104 control commands through wireless module. The control commands contain the helicopter's flight path. A servo controller board which is to control the switch between the automatic mode and manual mode. It not only guarantee flight safety but also ensure that the control algorithm to be tested and validated [9].

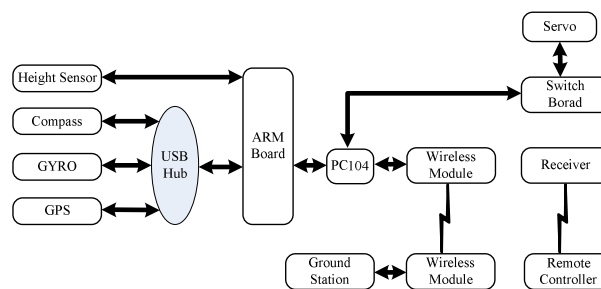


Figure 3: The logical connection diagram of rotorcraft UAV system.

3.1 Flight Control Computer Component

The computer system installed on the helicopter is PCM-3586, a typical industrial embedded computer system. PCM-3586 is a low power consumption ($4W@800MHZ$), X86 embedded motherboard which is designed specifically for PC104 field. The CPU type of PCM-3586 is SOC Vortex86DX, it has serial/parallel port, high-speed USB2.0, 10M/100M Ethernet, 256MB DDR2 system memory, 4G hard disk, 16 channel PWM output.

The installed operating system in PC104 is Fedora11, which integrates the GCC compiler and eclipse development environment. The main tasks of PC104 include:

- To maintain wireless communication with the ground station computer, the exchange data consist of control commands and monitoring information.
- Receiving sensor flight data, including flight attitude, altitude, heading, latitude and longitude information.
- To execute flight control laws with a main frequency of 50HZ, and generate PWM signal to drive actuators.

3.2 Flight Data Collector Component

The ARM9 industrial embedded computer system is used as Flight data collector. It has a Samsung S3C2440A processor, fully compatible with Linux OS, with a main frequency 400MHz and a 64MB SDRAM. One 256MB compact NAND flash disk is taken as the storage media of the Linux OS, which kernel version is 2.6.32.2 and integrated the Arm-Linux-GCC cross compiler environment. 100M Ethernet RJ-45 interface (using DM9000 network chip).

The flight data collector receive the sensors data includes: height(altimeter), flight heading(three-axis compass), the aircraft attitude(gyro), position (GPS), and then use the tracking differentiator filter way to isolate the biases. The output sensor data is transmitted to the flight control computer through the Ethernet which is based on UDP protocol.

3.3 Position Sensing Component

HC12 OEM can tracks GPS L2 and SBAS satellite signal produced by a Canada company. GPS provides three dimensional coordinate, which difference accuracy less than 0.5 meters. The GPS provides position estimates at 10Hz. The OEM board should connect with the GPS antenna to get the position information.

3.4 Attitude Sensing Component

The CS-VG-02 vertical gyroscope is used as the aircraft attitude sensor, which is inertial sensor based MEMS technology. The gyro is used to measure inclination of the helicopter relative to the horizontal plane, it has two sensitive axes respectively detect roll and pitch angle change, can output angle and angular rate signal simultaneously. The gyro angle static accuracy precision of less than 0.6 and dynamic accuracy of less than 2.5 degrees. It provides fast response time up to 200Hz and communication interface is RS422.

3.5 Heading Sensing Component

Flight heading sensor use the MWC3000L high-precision digital three-axis compass, which integrated three-axis magnetic resistance sensor and two-axis tilt sensor on-chip. It provides roll angle, pitch angle, heading angle and magnetic state, orientation repeatability of 0.5 degrees and inclination repeat of 0.1 degrees. The information update rate is 20HZ.

But we found that the heading angle signal generated by compass under the condition of vibration can not be used. Because of the roll and pitch angle in vibration condition is unstable, which leads to the heading angle instability. Accurate yaw angle can be calculated by the original data of three axes compass. Under the premise of making the compass's installation position parallel to the earth's surface, yaw angle can be calculated from the magnetic field in the X, Y direction components. The yaw angle ϕ calculation formula is shown as follows:

$$\phi = \begin{cases} 90 + \arctan(h_x/h_y) \times 180/\pi, & H_y > 0 \\ 270 + \arctan(h_x/h_y) \times 180/\pi, & H_y < 0 \\ 180, & H_y = 0, H_x < 0 \\ 0, & H_y = 0, H_x > 0 \end{cases} \quad (1)$$

where H_x is the X-axis direction of the magnetic component, and H_y is the Y-axis direction of the magnetic component. Direction is the heading signal of helicopter.

3.6 Low Height Sensing Component

In order to ensure that the aircraft can be controlled stability during the takeoff and landing state, the vertical height should be guarantee to be accuracy. In the take-off and landing conditions, height of the helicopter generally less than 1 meter in height control. Based on the above considerations, we selected the ultrasonic ranging module to complete the height measurement.

Ultrasonic ranging module uses the voltage DC5V, with less than 15 degrees sensitive angle, detecting distance is 2cm-450cm, accuracy is 0.3mm+1%, and the data update rate is 40HZ.

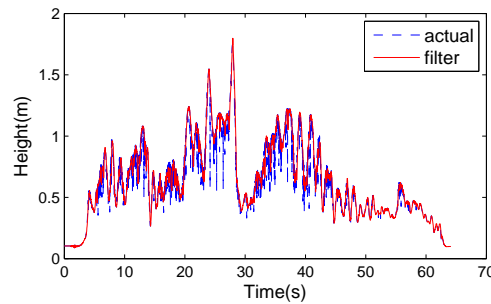


Figure 4: Height before and after TD.

The aircraft itself will generate about 30HZ vibration in the actual flight, which will affect the data of ultrasonic range module. The vibration can make the data we collected would emerge burr. We use tracking differentiator filter to filtering and tracking the signal. Fig.4 shows the effect before and after the filtering.

3.7 Actuation Component

The servo controller board has the function that switch the helicopter control between manual mode and automatic mode. Channel 7 in the radio control system is in charge of the mode of switch. The servo controller board shares a power supply with the servos. Such a design can ensure that the manual operation can run independently. The UAV helicopter can be operated in either the automatic mode or manual mode.

All of the servos are driven by a pulse width modulated(PWM) signal. So the input and output signal of the servo controller board are PWM signal. We use 74LS157 as the core selection chip, which is a type of either-or chip. The analysis of PWM signal is accomplished by a Micro control unit (MCU) C8051F330. PWM signal is resolved into high-low level, and then drive the selection chip. Each signal which in and out the servos controller board would be processed with light-coupled isolation, which can effectively prevent the signal interference.

The update rate of all sensors is ranging from 10-200HZ, which is enough for implementation for advanced control algorithms.

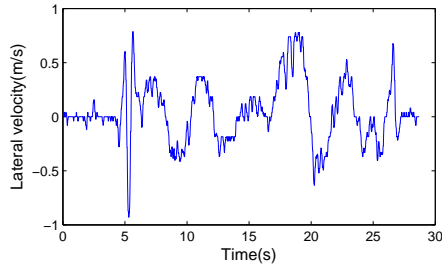


Figure 5: Lateral velocity.

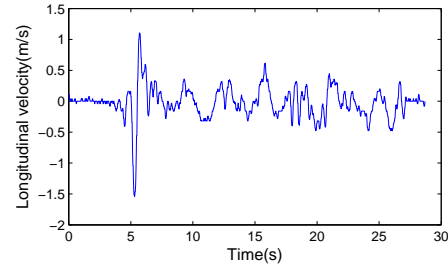


Figure 6: Forward velocity.

4 Cyber Algorithm Design

4.1 Online Flight Data Processing

GPS provides world geographic coordinate information, which can be applied to the position control algorithm. Gyroscope provides aircraft attitude information, which can be applied to the attitude control algorithm. But there is no sensor to acquire the velocity information of helicopter which need to be used in the velocity control algorithm. In order to get the velocity information, we adopt the tracking differentiator to tracking the location information of the GPS.

Tracking differentiator can extract reasonable differential signal of the measuring data with the predicted and tracked effect [10]. The general problem describe as a mechanical system that initial with the signal of position and velocity at t moment and output the signal of position and velocity at $t+1$ moment. The second order discrete forms of tracking differentiator is shown as follows:

$$\begin{aligned} x_1(t+1) &= x_1(t) + h \times x_2(t) \\ x_2(t+1) &= x_2(t) + h \times \text{fhan}(x_1(t) - x(t), x_2(t), r, h_0) \end{aligned} \tag{2}$$

where x_1 describing the motion of the target's position and x_2 is the velocity information of the target. h is the sampling step. The function $\text{fhan}()$ is shown as follows:

$$\text{fhan} = \begin{cases} -r \times a_1/d, & \|a_1\| \leq d \\ -r \times \text{sign}(a_1), & \|a_1\| > d \end{cases} \tag{3}$$

with

$$\begin{aligned} d &= rh_0 \\ d_0 &= dh_0 \\ y &= x_1 - x + h_0x_2 \\ a_0 &= \sqrt{d^2 + 8r\|y\|} \\ a_1 &= \begin{cases} x_2 + y/h_0, & \|y\| \leq d_0 \\ x_2 + \text{sign}(y)(a_0 - d)/2, & \|y\| > d_0 \end{cases} \end{aligned} \tag{4}$$

where r is speed factor and h_0 is the filter factor, speed factor determines the tracking velocity, filter factor determines the filter effect. The filter effect is proportional to the h_0 , and the level of phase lose is inversely proportional to the h_0 .

The input signal of the tracking differentiator is velocity of helicopter which is calculated by the GPS information. And its output signal is the predicted velocity of helicopter. The effect of forward and lateral velocities is given in Fig.5 and Fig.6.

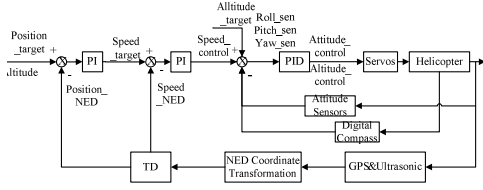


Figure 7: Position controller scheme.

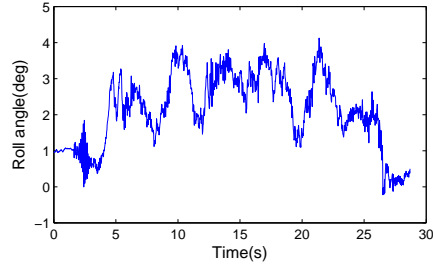


Figure 8: Roll angle during the flight

4.2 Control Algorithm Design

The flight control system algorithm uses the body coordinates as the navigation coordinate. The position information which GPS provides is in the WGS-84 coordinates. First of all, the information of GPS (longitude, latitude, altitude) need to be converted into world geographic coordinates (NED coordinates), and use the position data when helicopter take-off as the original point. Then we must transform the NED coordinates into the body coordinates during the navigation control process [11].

Simulation studies have shown that a better strategy for the control of a small-scaled helicopter is to use the flight controller as consisting of three cascaded controllers: an inner loop attitude control, a middle loop velocity control and an outer loop position control [12]. The overall flight control scheme is shown in Fig.7.

Height Control

The height control is a one loop scheme which is a PI controller using feedback from ultrasonic range finder generates collective pitch demands. In order to improve the control stability, we use the vertical velocity signal as the compensation for the collective pitch demands. The calculated formula is shown as follows:

$$\begin{aligned}
 Z_e &= z_{current} - z_c \\
 V_{zc} &= K_p^{zv} Z_e \\
 V_{zc} &= [V_{zc}]_{-V_z^{lim}}^{V_z^{lim}} \\
 Z_s &= (V_z - V_{zc}) \times K_p^{zv} \\
 A_c &= K_p^z Z_e + K_i^z \int_0^t (Z_e + Z_s) dt
 \end{aligned} \tag{5}$$

where $z_{current}$ is the actual height of helicopter and z_c is the target height. Z_e is the height error and it can be used to calculate the Z_s which is the compensation for the collective pitch control signal. A_c is the collective pitch control signal. V_z^{lim} is the limitation of the speed of hight. p^{zv} and K_p^z are the proportional control parameters. K_i^z is the integral control parameter.

Yaw Control

The yaw control is a proportional controller using the feedback from the digital compass generates the rudder channel control demands. GY401 is used to increase yaw control stability. According to body sway velocity, GY401 will make reverse compensation signal to the actuator and suppress the helicopter tail rotor's swing which just as a tail rotor damper.

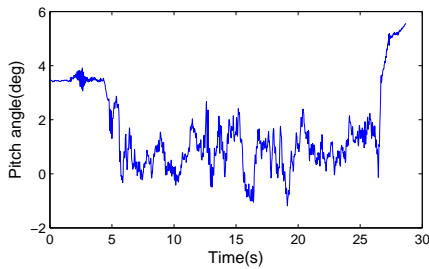


Figure 9: Pitch angle during the flight.

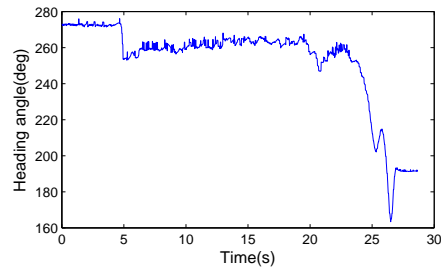


Figure 10: Heading angle during the flight.

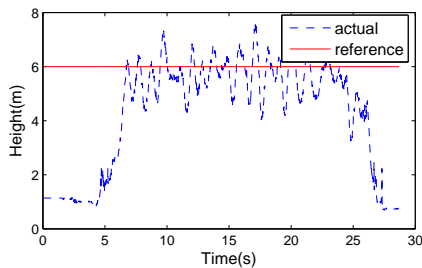


Figure 11: Height control during the flight.

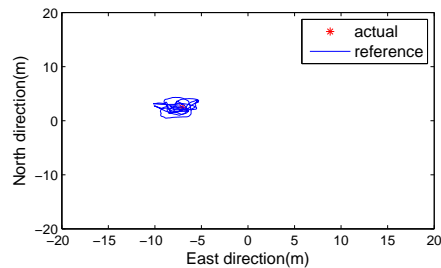


Figure 12: Position control during the flight.

Position Control

The position control is a three-cascade scheme which includes an inner loop, a middle loop and an outer loop. The outer loop takes target position as desired position as inputs and generates desired velocity to the middle loop which using the feedback from the GPS. The middle loop is designed as velocity controller which takes desired velocity as inputs and generates the desired attitude angles to the inner loop. The inner loop is designed as the attitude controller which takes desired attitude angles as inputs and generates the actuator commands that drive the helicopter reach the desired attitude. Finally the helicopter will reach to the target position through the control demands.

5 Real-Life Autonomous Flight Experiments

A three-loop control scheme for rotorcraft UAV system was design and implemented using TREX 600 platform. We use the fixed-point hovering experiment to show the controlled effect. The target point is at longitude 12126.31838 and latitude 3101.44161, which change into the NED coordinates is XYZ(-7.0658, 2.4706, 6). X stand for the west-east coordinate, Y stand for the south-north coordinate, and Z is the height coordinate. Fig.8-Fig.10 show three-axes angles during the flight. Fig.11 and Fig.12 show that position which controlled by the outer loop is get a stable response. The control accuracy is of a radius of two meters.

6 Conclusion

This paper describes the Cyber-Physical System implementation of the rotorcraft UAV and control scheme based on TREX600 aeromodelling helicopter. The rotorcraft is a small helicopter model, which will be changed to adapt to the heavy load in future. The rotorcraft UAV system has been tested successfully for automatic flight, including take-off and landing.

The system also has some deficiencies, such as poor wind-resistant performance. There is an influence on the control precision when the wind speed is 5m/s. The flight path is not stable enough, and the altitude of helicopter would be effected when the helicopter flight forward. The next step is to integrate the wind model into the existing control model so as to improve the flight stability, and add new application tasks.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (No.61303013), Shanghai Natural Science Foundation (No.12ZR1445700) and Research Fund for the Doctoral Program of Higher Education of China (RFDP) (No.20120073120039). Note: The authors contribute equally for the work in this paper.

Bibliography

- [1] Yao, J. et al; NetSimplex: Controller Fault Tolerance Architecture in Networked Control Systems, *IEEE Transactions on Industrial Informatics*, 9(1), 346-356, 2013.
- [2] Nonami K; Prospect and recent research & development for civil use autonomous unmanned aircraft as UAV and MAV, *Journal of System Design and Dynamics*, 1(2),120-128, 2007.
- [3] Johnson, E. et al; The georgia tech unmanned aerial research vehicle GTMax, *Proc. of AIAA Guidance, Navigation, and Control Conference*, Austin, 2003.
- [4] Qi, J. et al; The servoHeli-20 rotorcraft UAV project, *Proc. of the 15th Int. Conf. on Mechatronics and Machine Vision in Practice (M2VIP 08)*, Auckland, New-Zealand, 2008.
- [5] Cai, G. et al; An overview on development of miniature unmanned rotorcraft systems, *Frontiers of Electrical and Electronic Engineering in China*, 5(1), 1-14, 2010.
- [6] Zhou, W. et al; Research on UAV Flight Control Computer Based on ARM, *Computer Measurement & Control*, 17(7),1286-1288, 2009.
- [7] Cai, G. et al; Construction, modeling and control of a mini autonomous UAV helicopter, *Proc. of IEEE Int. Conf. on Automation and Logistics (ICAL 08)*, Qingdao, China, 2008.
- [8] Gu, Y. et al; Integrated avionics system for research UAVs, *Proc. of AIAA Guidance Navigation and Controls Conference and Exhibit*, Honolulu, Hawaii, 2008.
- [9] Cai, G. et al; Development of fully functional miniature unmanned rotorcraft systems, *Proc. of the 29th Chinese Control Conference (CCC 10)*. Beijing, China, 2010.
- [10] Su, X. et al; The research on single-antenna DGPS determination attitude method based on tracking-differentiator, *J. of Projectiles, Rockets, Missiles and Guidance*, 31(2), 199-204, 2011.
- [11] Peng,K. et al; Design and implementation of a fully autonomous flight control system for a UAV helicopter, *Proc. of the 26th Chinese Control Conference*, Zhangjiajie, China, 2007.
- [12] Shim,D. et al; Hierarchical Control system synthesis for rotorcraft-based unmanned aerial vehicles, *Proc. of AIAA Guidance Navigation and Controls Conference and Exhibit*, Denver, Co, Aug, 2000.

Distributed Genetic Algorithm for Disaster Relief Planning

K. Zidi, F. Mguis, P. Borne, K. Ghedira

Kamel Zidi

Science Faculty of Gafsa, university of Gafsa
Campus Universitaire Sidi Ahmed Zarrouk 2112 Gafsa, Tunisia
kamel_zidi@yahoo.fr

Fethi Mguis*, Khaled Ghedira

Higher Management School of Tunis, University of Tunis
41, Rue de la liberté Bouchoucha le Bardo 2000, Tunisia
*Corresponding author: fethi.mguis@fsg.rnu.tn
khaled.ghedira@isg.rnu.tn

Pierre Borne

Ecole Centrale de Lille
Villeneuve d'ascq 59650, France
pierre.borne@ec-lille.fr

Abstract: The problem studied in this paper is the management of vehicle routing in case of emergency. It is decomposed into two parts. The first one deals with the emergency planning in the event of receiving a set of requests for help after a major disaster such as in the case of an earthquake, hurricane, flood, etc. The second part concerns the treatment of contingency as the arrival of a new request or the appearance of a disturbance such as breakdowns of vehicles, the malfunction of roads, availability of airports, etc. To solve this problem we proposed a multi-agents approach using a guided genetic algorithm for scheduling vehicle routing and local search for the management of contingencies. The main objective of our approach was to maximize the number of saved people and minimize the costs of the rescue operation. This approach was tested with the modified Solomon benchmarks and gave good results.

Keywords: Vehicle Routing Problem, multi-agent system, genetic algorithm, emergency, disaster relief.

1 Introduction

Although there were several studies in the field of vehicle dynamics deals with the problem of the distribution (eg, [9]- [34]- [35]- [39]). However, most of the available research has focused on commercial transport, logistics planning in supply chain repetitive, and routine treatments and environments orders [4]. Studies dealing with vehicle routing problems as the dynamic response planning and logistics of disaster relief were few.

The objective of this work was to propose a distributed approach based on a genetic algorithm for modeling and solving the problem of contingency planning in case of disaster. Our approach enables a planning disaster relief distribution by maximizing the number of people saved, avoiding delays in the relief distribution and maximizing the useful equipment life which saves costs.

In section 2, we have presented the problems and the aims of this work; while section 3 presented a literature overview dealing with the problem of emergency planning for disaster. In the section four highlighted and described the proposed approach. In the last two sections, we have discussed tests of calculation and managerial implications arising from the study, and we have summarized the work, by presenting findings and suggesting further alternatives.

2 Problematic and Aims

As the world continue to witness the vulnerability of natural and artificial disasters on human well-being is constantly under threat (eg, [3] [22] [36] [37]). For example, about two million people were affected by the Haitian earthquake of January 2010, the Haitian government have reported that an entire nation of 230,000 people had died, 300,000 were injured and 1,000,000 made homeless [47]. Furthermore about 250,000 homes and 30,000 commercial buildings were collapsed or were severely damaged [47]. Similarly, the earthquake of 2008 in China killed more than 67,000 people. In 2008 Cyclone Nargis killed more than 78,000 in Myanmar [28], the Kashmir earthquake in Pakistan in 2005 killed more than 86,000 people and Katrina Hurricane killed more than 1,000 person with damages of billions of dollars [21].

In fact, Disasters and crises are often characterized by massive displacement of people, lack of food and water, degradation of essential services, damage to infrastructure, unsatisfactory warehousing capacity, inadequate transportation and inaccessibility to remote areas [28] [29]. Add to this, other challenges include the difficulty in assessing the needs and expectations of victims (human resources management), insufficient relief or quantity transported and delivered as well as issues of inter-agency coordination.

Often, disaster response was characterized by excessive centralization and short-term nature of the emergency chain coupled with a lack of reliable information [1] [18] [28]. In addition, sources the supply distribution nodes, and extended distribution points must be quickly putted in place, sometimes because of the topographical difficult situations of pickups and deliveries [6] [19]. In several times, organizations and agencies have negotiated with governments and administrations, military access, municipal authorities and organizations [4] [33], therefore, it appeared that crises and disasters create circumstances and extraordinary conditions for those who seek to manage relief and rescue operations. These actors must take urgent decisions while essential information on the causes. Thus, consequences remain unavailable and the degree of uncertainty is high [38]. Despite this difficult environment disaster with rapid change, and the efforts of charities, aid agencies and governments, stakeholders such as emergency beneficiaries [12] [23] [32] as well as the media [8] [48] have criticized how the relief was distributed and the disaster relief chains were managed. It's the case of inefficiency, waste, and lack of planning and logistical capacity limited disaster were cited. These criticisms, were also applied to the disaster relief operations, for example, in the aftermath of Hurricane Katrina in 2005.

Such an investigation was crucial for improving real-time response and effectively to the consequences of disasters as well. Other crises where relief goods such as water, food, first aid and so must be rapidly distributed to reduce human suffering and save lives. In others words, the nature of a disaster, especially with regard to the availability of infrastructure logistics, warning, scale, location, topography and resources have an impact on the design plans and emergency response, as well as the mixture of the allocation of resources [33].

It appeared that, some parameters must be taken into consideration to better understand the situation of the typical distribution of disaster relief. Resources of the fleet of vehicles and emergency equipment were generally insufficient; the affected area was often a poor logistics infrastructure (eg, bad and damaged roads , warehouses, ports, etc.) and desperate people who may be dispersed to different nodes (feeding sites and/or distribution) over a large area. Therefore, it is difficult for both relief workers and logisticians to rapidly develop effective and efficient distribution plans [4] [34] [35].

Second, fleet management with interest on routing, planning, pickup and delivery are often the main decisions to be taken by the logisticians and the relief's effective allocation and satisfaction of requests place were also to be taken in consideration. These decisions are important for the distribution of relief because of "life or death." This requires fast delivery with strict limits

required by donors and financial accountability reports. Hence the need for an efficient routing and planning essential to the effectiveness of relief programs.

3 Literature Review

In recent years, the problems of disaster emergency management have attracted more and more the researchers's attention. This was due to their importance for human, economy and society. Indeed, some approaches have been proposed for modeling and solving this problem. Oh and Haghani [27] modeled the problem as a linear program by considering only how minimizing the cost. However, Tzeng [41] developed a heuristic based on linear program in which they took into account the human and the cost; while Yi [42] proposed an approach based on ACO meta-heuristic modeling with linear program witch consider only the human side of the problem. in same way Ozdamar [30] proposed a modeling of the problem as a linear program. Furthermore, he also proposed a modeling of the problem as a linear program in which they took into account the human side and they have developed a heuristic to solve the problem [31]. In contrast, Ma [20] proposed an exact method using a nonlinear program by considering the human goal only. This was similar to some extend, Afshar [2] is based on an exact method to solve the problem by modeling as a linear program that takes into account the human side. Zhang [44] developed a heuristic which considers the minimization of cost.

To conclude, we can say that:

- Most approaches do not take into account the aspect of multi-objective problem.
- The majority of available studies dealing with static problems ignore the dynamic aspect of the problem.
- The meta-heuristics were less used despite they were more efficient especially with this kind of problem since they require a short response time with a number of exponential data.
- All the approaches developed did not include forecasts to predict future requests for help. In fact, in the disaster case, forecasting future demands have allowed to avoid certain circumstances and reduce the damage to human and material.
- Distributed approaches were almost absent despite their great performance for modeling and solving problems similar to the problem addressed herein(eg, [25]- [24]- [45]- [43]- [7]- [17]- [10]- [14]).

4 Proposed Approach

In order to solve the above described problem, we propose a multi-agent model as developed by Zidi [46] to which we were add some features that we suspect necessary for the proper functioning of the system to adequately address the problem. Our model as described in Figure 1 was composed by a "System Agent" and a set of "Zone Subsystems". The "System Agent" has to create and supervise all other agents. Each "Zone Subsystems" was composed of a "Planning Subsystem", an "Area Manager Agent", "Information Manager Agent", "Forecast Agent", "Disturbance Agent", and a local database. The "Zone Subsystem" was managed by the "Area Manager Agent" which provides communication with the "System Agent" and other "Zone Subsystems". The "Information Manager Agent" was responsible for the management of information

necessary for the proper conduct of the rescue operation. In fact it provides information collection, their filtration, and updates the local database. The "Forecast Agent" allowed forecasting of emergency calls based on the information received and based on historical relief interventions. The "Disturbance Agent" will be used to detect disturbances that can occur, transmit "Planning Subsystem" to be processed. The "Planning Subsystem" was responsible for emergency planning, integration of new applications and correction planning in case of a disturbance. The "Planning Subsystem" was composed of three types of agents namely "Supervisor Agent" who supervises the other components in addition to its roles in the "Planning Subsystem". The role of "Vehicle Agent" is finding a path for a vehicle. The "Interface Agent" handles communication with the external environment. The local database of "Zone Subsystem" will be used to store all data in the backup area. Each "Zone Subsystem" can communicate with the external environment mainly composed by: applicants emergency disaster, relief agencies, the geographic information system GIS, news and weather agencies.

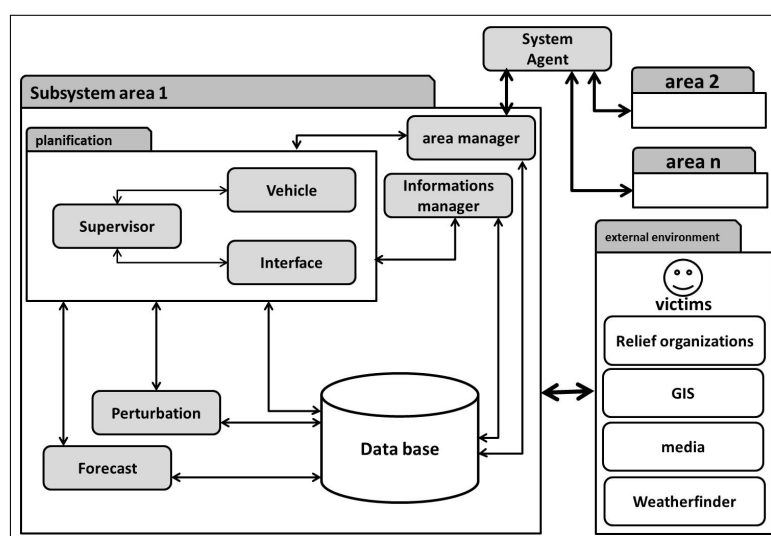


Figure 1: Overview of the system

4.1 The Planning Subsystem

The subsystem of planning presented in Figure 2 was composed of three types of agents: Supervisor, Interface and Vehicle.

Supervisor Agent

This agent ensures the consolidation of emergency calls, the creation of tenders for the insertion of a new application, choose the best solution and update planning after any changes generated by the insertion of a new application or by the occurrence of a disturbance.

Interface Agent

This agent was responsible for all information exchanged between the "Planning Subsystem" and its external environment. It was responsible for receiving requests for help formatting, negotiating with transport operators and information from stakeholders.

Vehicle Agent

Each vehicle agent was responsible for a vehicle. It has to ensure that the search path will be followed by the vehicle. In addition to receiving a bid for the insertion of a new request, the agent "Vehicle" checks can add this request and sets its offer, which sends it to the "Supervisor Agent".

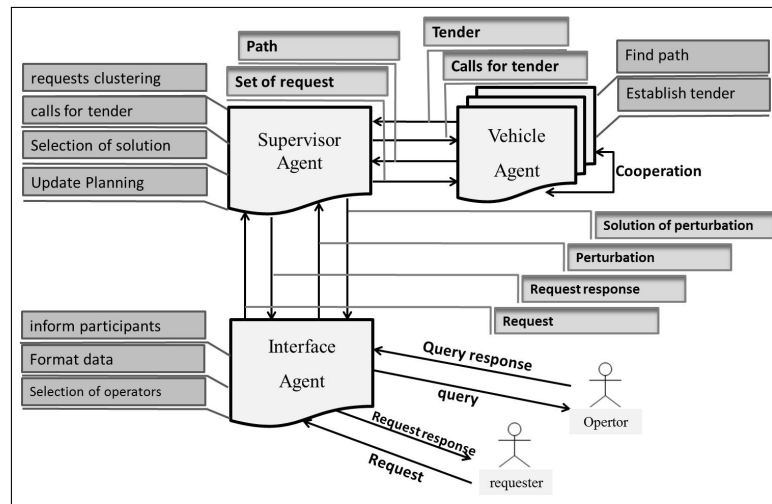


Figure 2: Subsystem of planning

Agents Communications

We identify ten types of communication messages between agents of "Planning Subsystem". These messages were: "Request" was sent by the "Interface Agent" to the agent "Supervisor" upon receipt of a new request for help. After processing a new application, a "Request Response" was sent by the "Supervisor Agent" to the "Interface Agent". When receiving an announcement of a disturbance, a message "Disturbance" has to be sent by the "Interface Agent" to the "Supervisor Agent". After treatment of the disturbance, a message "Solution of disturbance" which was sent by the "Supervisor Agent" to "Interface Agent". After the consolidation of emergency requests, messages "Group requests" were sent by the "Supervisor Agent" to "Vehicle Agents". After searching for a path, a message "Path" was sent by the "Vehicle Agent" to the "Supervisor Agent". Following the receipt of a new request messages "Tender" were sent by the "Supervisor Agent" to "Vehicle Agents". After the establishment of an offer to insert a new request, a message "Offer" was sent by a "Vehicle Agent" to "Supervisor Agent". In order to improve the quality of its way, a "Vehicle Agent" can send a message "Negotiation" to the other "Vehicle Agent". Communications between the different agents were presented in Figure 3.

Emergency Planning

Emergency planning starts by grouping the applications according to their types, their locations, their degrees of urgency, their quantities, and their time windows. Then each group will be assigned to an "Vehicle Agent" which was responsible for finding the right path according to a heuristic which sort the requests in ascending order according to their service termination dates. In the of equality case the customer closest to the preceding customer has to be selected. This creates an initial population for a genetic algorithm executed by the "Supervisor Agent" altogether with the "vehicle Agent". At the end a we got the final population of the solutions

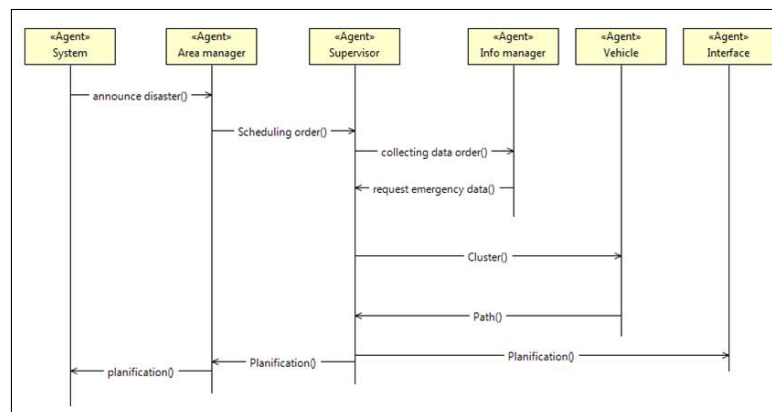


Figure 3: Protocols diagram for planning

from which the "Supervisor Agent" have to chose one of them to be followed for the planning of the rescue operation.

Genetic Algorithm

The real universal problems, including disaster cases has inevitably some unknown and uncertain parameters. Stochastic programming seems to be the most appropriate to solve this problems, while it requires probability distributions for the estimate data. The solution computed should be valid for the majority of possible realities and the expected value of the objective function was optimized. Our challenge was to propose an alternative approach to provide probability distributions. If the deviations were within certain limits, our idea was to find the valid solution, even if small changes may occurs. Therefore, the desired visits should be robust and flexible. However, these terms were often used with different meanings; hence we will define this terms to be used later in this document:

- Robustness: if a change in travel time or a new customer request arrives, only minor changes were needed in the planning.
- Flexibility: it will keep the schedule, because the generated plan has several options.

Therefore, a right option was to solve a stochastic model that was particularly difficult to solve in a certain period. Alternatively, it is possible to solve a deterministic model with additional constraints to create a robust and similar flexibility. Therefore, a similar performance with less efforts calculation can be performed. So with such problem, the approximate methods have proved more effective resolution. In fact, they allow to obtain acceptable solutions in a reasonable time. Among the approximate methods the most used wre the genetic algorithms that provide solutions which guaranteed robustness and flexibility.

The fundamental principles of these algorithms have been incurred by Holland [15]. These algorithms are based on the functioning of the natural evolution of species, including the selection of Darwin and Mendel procreation. They have been successfully used to solve several problems of multi-criteria optimization [11].

In genetic algorithms, we simulate the process of evolution of a population. One starts with an initial population composed of N solutions (individuals) of the problem. The degree of adaptation of an individual to the environment is expressed by the value of the cost function $f(x)$, where x is the solution that is the individual. It is said that a person is much better adapted to its environment, the cost of the solution is lower or larger depending on test selected optimization.

Within this population, occurs when the random selection of one or both parents, producing a new solution, through genetic operators such as crossover and mutation. The new population is obtained by the choice of N individuals among populations (parents and children) is called next generation. By iterating the process, we produce population richer individuals best suited. The process of the genetic algorithm is presented in Figure 4. For success and the convergence of the

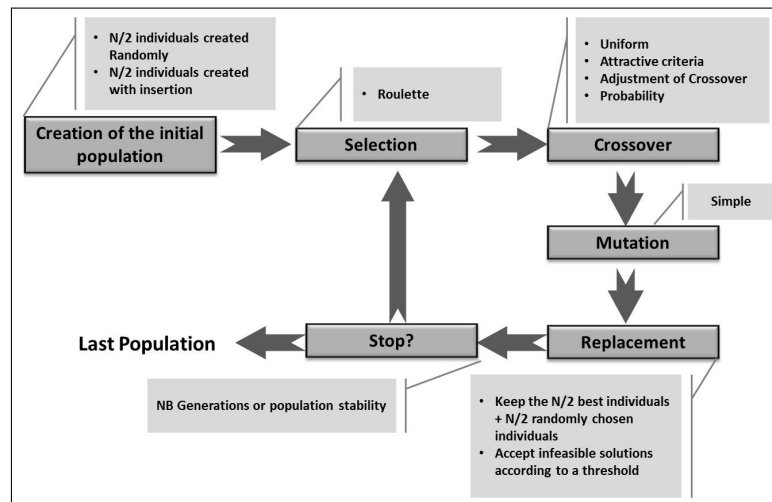


Figure 4: Functioning of the genetic algorithm

genetic algorithm requires:

- Make a good representation (encoding) of a solution (chromosome)
- choose the manufacturer of the original solution,
- clearly define the evaluation function to determine the fitness of each individual
- choose the methods of selection, crossover and mutation are best suited to the problem and
- carefully choose the parameter values: population size, probabilities, etc..

Coding: Because the chromosomes are not binary adopted for sequencing problems, it has adopted an actual coding where the chromosome is a sequence of nodes (excluding the deposit). Figure 5 shows a chromosome of a problem consists of 10 clients.

Creation of the initial population: To build the initial population, we have developed a heuristic which builds the first half of the population randomly and the second half in a guided manner based on the method of insertion of Solomon [40].

Selection: To select the chromosomes (solutions) that will be able to contribute to the creation of the new population, we adopted the method of selection of Wheel [15] [13] which consists in assigning to each individual a probability of selection proportional to its evaluation (fitness) and the sum of the evaluations of individuals. The selection process is shown by the following algorithm:

1. Calculating the fitness f_i for each individual of the population

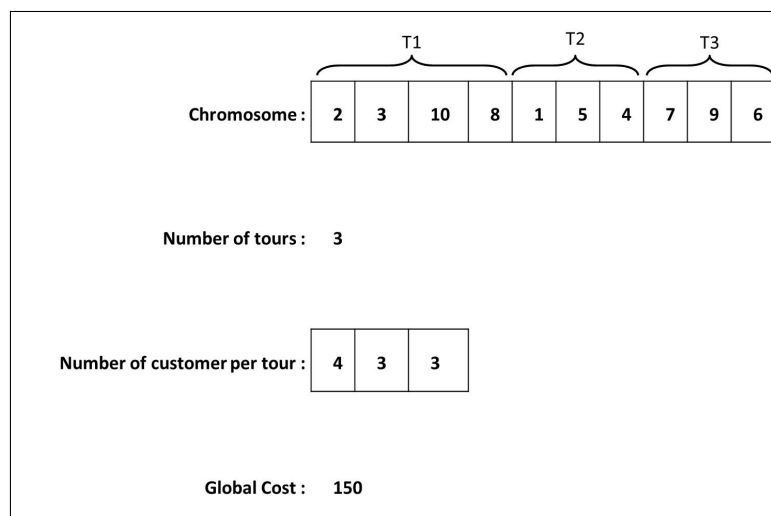


Figure 5: Example of an individual

2. Calculating the probability P_i of selection of each individual $P_i = f_i / \sum_{i=1}^{i=n} f_i$
3. Calculation of cumulative probability q_i of each individual $q_i = \sum_{j=1}^i p_j$
4. Generate a random value $r \in [0, 1]$
5. If $r < q_1$ then select the first individual else choose individual i as: $q_{i-1} < r \leq q_i$
6. Repeat steps 4 and 5 to create the desired number of individual

Crossover: From both parents (solutions), we try to generate one or two son. There are several breeding techniques. For each type of problem, there are a set of breeding methods that are more suitable. For the vehicle routing problem, according to the literature, the most commonly used methods are: a crossing point, dual point crossover and uniform crossover. In our case we used the Uniform crossover. This method consists in determining the cost, the number of customers and the ratio of these last two for each round of two parents chosen at random. Then, each parent towers are sorted in ascending order according to the quotient: Cost/Number of customers. The third step is to establish a conflict matrix to describe the conflicts between the towers of both parents. The construction of the individual descending begins by inserting the first turn of one of the parents. The second step in the construction of the son is to exclude all rounds of the other parent are in conflict with the tour recently added. The same process is then repeated with the first round of the other parent and so on until there are more towers to insert. At this stage it is possible to customers not served. These customers are inserted into the shot so that the overall cost is minimal. Figure 6 illustrate an example of a crossover operation.

Mutation: In the process of evolution, mutation performs a broader exploration of the search space, to avoid premature convergence and diversity loss by bringing innovation in the population. In our approach we have adopted the method of simple random mutation. The principle of this method is to choose two nodes randomly. Result in generating a random value between 0 and 1. If this value is less than the mutation probability, then swap the two nodes, otherwise do nothing. Figure 7 shows an example of a mutation operation.

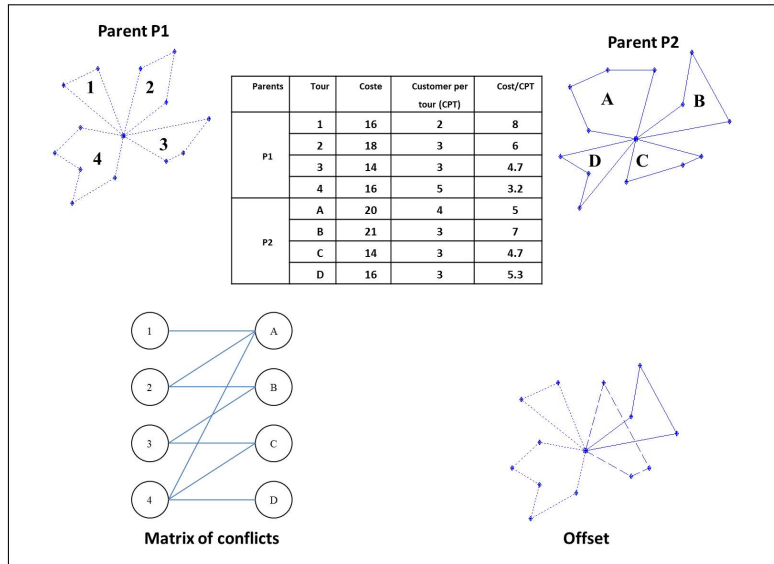


Figure 6: Example of a crossover operation

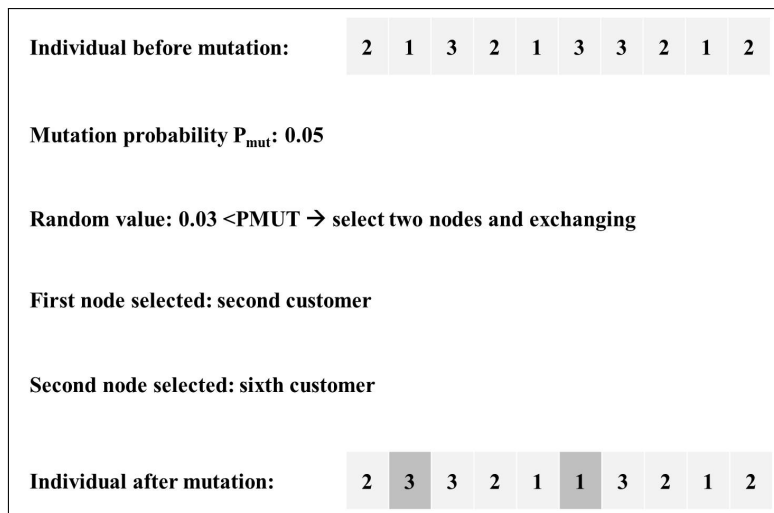


Figure 7: Example of a mutation operation

Replacement: After each iteration, individuals created (children) will be added to the population. To keep the same population size, we proceed to an alternative transaction elitist is to sort individuals according to their overall costs and keeps only the first N individuals to form the next generation.

4.2 Treatment of a New Request

Upon the arrival of a new request for assistance to the agent "Interface", it begins with collecting the necessary information that can help the agent "Supervisor" to make the right decision, then it sends the request to the "Supervisor Agent". This establishes a tender which will be sent to "Vehicle Agents" which will be able to meet this demand. Each "Vehicle Agent" receiving the tender should calculate its area of action [43][60] which is equal to the number of customers (t, x, y) that can be served by this vehicle (t_0, x_0, y_0) . Then this agent should calculate also the insertion cost which equals to the difference between the old and the new zone area. Finally, the "Vehicle Agent" sends its offer to the "Supervisor Agent" who is in charge of choosing the best offer and sends the response to the "Interface Agent" which should inform stakeholders involved in this new application. If the "Supervisor Agent" could not insert the new request in any tour, there will be creation of a new tour to meet request. The above process of a new request treatment was described in Figure 8.

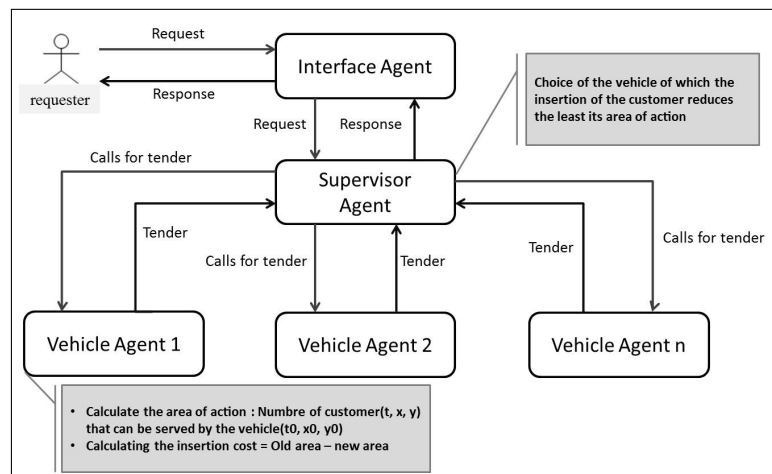


Figure 8: Treatment of a new request

4.3 Treatment of a Disturbance

A good management of disaster relief should take into account the occurrence of disturbances at any time during the rescue. In fact in the case of disasters road and emergency equipment were very delicate and it may cause a malfunction of the system. For these reasons, we have provided a module for the treatment of disturbances. When receiving an announcement of a disturbance, the "Interface Agent" was responsible for the collection of information relating to the obstruction. Then it sends all this information to the "Supervisor Agent" who takes care of looking for a solution to the disturbance. Disturbances treated herein were two types. Disturbance may be caused by a vehicle breakdown or unavailability of a road. In the vehicle breakdown case, the "Supervisor Agent" checks the possibility of replacing the vehicle. if not he have to try to fit the request of the disabled vehicle in the other tour, without interfering any constraints. In case of unavailability of a road, we proposed two alternatives. The first one

consist of changing the vehicle way by avoiding the disrupted roads. The second alternative consist of inserting customers in other tours. The treatment process of a disturbance was shown in Figure 9.

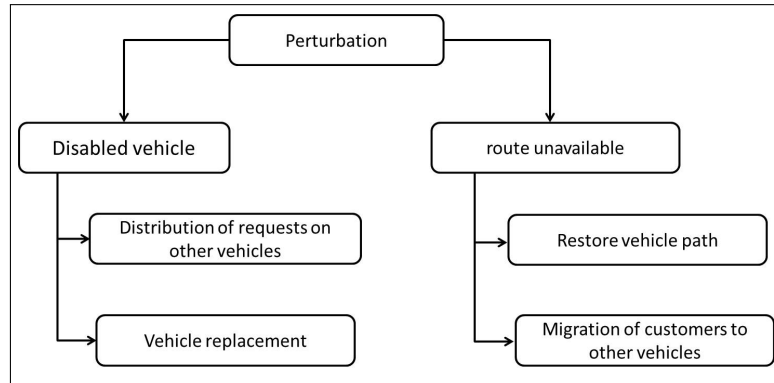


Figure 9: Treatment of a disturbance

5 Experiments and Results

In this section, we report our approach results in comparison with other algorithms. The tests data were described in the section 5.1; while the experimental results of the different problems were summarized in the section 5.2.

5.1 Experimental test data

The basic data of our testing problems adopt Solomon's 100-customer benchmark problems [40] for the static vehicle routing problem with time windows (VRPTW). In these benchmark problems, 100 nodes are distributed in a Euclidean plane of 100*100 squares, and the travel times between nodes are equal to the corresponding Euclidean distances. There are six types of problems, named C1, C2, R1, R2, RC1 and RC2, each with 8-12 problems. Different types of problems differ in the distribution of the nodes, the service time of each node, and the width of time windows, described in detail as follows:

- The locations of the nodes are clustered distribution in the problems of types C1 and C2, are random distribution in the problems of types R1 and R2, and are mixture distribution in the problems of Types RC1 and RC2.
- The service time at each node is 90 time units in the problems of Types C1 and C2 and 10 time units in the problems of types R1, R2, RC1, and RC2.
- The problems of types C1, R1 and RC1 have the vehicles of relatively small capacity; the capacity of each vehicle is 200 units. The ratios of the average demand of the nodes to vehicles capacities are 7.29%, 9.05% and 8.62%, respectively. And the problems of types C2, R2 and RC2 provide more capacity vehicles; the capacity of each vehicle is 700 units in Type C2, 1000 units in types R2 and RC2. The ratios of the average demands of the nodes to vehicles capacities are 1.469%, 2.59% and 1.72%, respectively.
- In each problem, there is a time window $[e_0, l_0]$ associated with the depot with in which a vehicle must return to the depot after serving some customers. The problems of Types

C1, R1 and RC1 have an arrow time window of the depots, such that only few customers can be covered in each trip, while problems of Types C2, R2 and RC2 have a wider time window of the depots.

5.2 Experimental Results

The test results were shown in Tables 1 for all the six type problems described in the section 4.1. For description convenience, we call the column (1) C1, column (2) C2, etc. NOV were the average of the used vehicles number. TD were the average of travel distance. C1, C2 were the computational results for the improved LNS algorithm (ILNS for short) which was proposed by Hong [16]. The C3 was the relative error between C2 and C8 (i.e. $C3=(C8-C2)/C2*100\%$). C4 and C5 are the computational results for the Two-stage hybrid local search approach (HLS for short) which was proposed by Bent and Van Hentenryck [5]. the C6 was the relative error between C5 and C8 (i.e. $C6=(C8-C5)/C5*100\%$). C7 and C8 were the computational results for our approach Guided Genetic Algorithm(GGA) [26]. C9 represents the percentage gain in execution time obtained by the use of our distributed approach. Based on these results, we can make the following conclusions on the effectiveness of the proposed approach to solve the problem studied:

- The solution quality generated by our approach can approximate to the best known solutions, the ILNS approach and the HLS approach. If the travel distances are only observed and the number of used vehicles were ignored, the average relative error of each type was less than 10%.
- Our approach provided an effective result within a reasonable time, which is very important in the disaster case. In fact, in the emergencies case, the solution should be provided quickly and efficiently. In addition our approach allowed at any time to provide an acceptable solution to stopping the execution of the algorithm after a given generation if necessary.
- Using multi-agent systems has allowed us to reduce the execution time which was very important for emergency planning in disaster case.

Table 1: Experimental results.

Problem type	ILNS approach			HLS approach			GGA		Our approach
	NOV (1)	TD (2)	ER(%) (3)	NOV (4)	TD (5)	ER(%) (6)	NOV (7)	TD (8)	Δ Time(%) (9)
C1	10	833,1	-0,53	10	828,82	-0,02	10	828,65	-25,46
C2	3	590,31	9,09	3	589,86	9,17	3	643,94	-22,03
R1	12,25	1218,28	-1,28	11,92	1244,52	-3,36	11,9	1202,74	-14,67
R2	3,27	964,11	-5,24	4	954,27	-4,26	2,9	913,6	-16,53
RC1	12,13	1369,57	0,46	11,5	1384,17	-0,60	11,9	1375,89	-17,89
RC2	3,75	1131,18	7,69	3,25	1124,46	8,34	3,9	1218,2	-28,91

6 Conclusions and Future Works

In this paper, we have developed a solution to the problem of emergency planning for disaster. We have illustrated some of the complexities practices for emergency in disaster case. Despite the importance of emergency planning in several disaster cases, they have received few attention

in the literature of the disaster and operations research. Thus, it appears that the extension of our approach to disaster relief, as described looks promising and with considerable value. Even if we know the boundaries explicitly construct models of reality and the unique nature of disasters, we suggest that operations research systems decision support can be beneficial in the disaster. Although several emergency relief organizations often hide their distribution planning for security reasons. finally we conclude that the adaptation and the use of the approach described above can contribute to the improvement of the use of the vehicle fleet management as well as the routing and delivery of relief goods in disaster cases.

Bibliography

- [1] Adinolfi, C. et al.; *Humanitarian response review, An independent report commissioned by the United Nations Emergency Relief Coordinator & Under-Secretary-General for Humanitarian Affairs*, Office for the Coordination of Humanitarian Affairs (OCHA), New York and Geneva, 2005.
- [2] Afshar, A.; Haghani, A.; Modeling integrated supply chain logistics in real-time large-scale disaster relief operations, *Socio-Economic Planning Sciences*, 46(4): 327-338, 2012.
- [3] Baer, M. et al.; *Safe: the race to protect ourselves in a newly dangerous world*, New York: HarperCollins, 2005.
- [4] Balcik, B. et al.; Last mile distribution in humanitarian relief, *Journal of Intelligent Transportation Systems*, 12(2): 51-63, 2008.
- [5] Bent, R.; Hentenryck, P.V.; Scenario-based planning for partially dynamic vehicle routing with stochastic customers, *Operations Research*, 52(6): 977-987, 2004.
- [6] Beresford, A.; Rugamba, A.; *Evaluation of the transport Sector in Rwanda*, Geneva: UNCTAD, 1996.
- [7] Boudali, I. et al.; An Interactive Distributed Approach for the VRP with Time Windows, *Journal of Simulation Systems, Science and Technology*, 2005.
- [8] Burg, S.; Shoup, P. (1999); *The war in BosniaHerzegovina: ethnic conflict and international intervention*, Armonk, NY: M.E Sharpe, 1999.
- [9] Campbell, A.M.; Savelsbergh M.W.P.; A decomposition approach for the inventory-routing problem, *Transportation Science*, 38: 488-502, 2004.
- [10] Cheng, C.B.; Wang, K.P.; Solving a vehicle routing problem with time windows by a decomposition technique and a genetic algorithm, *Expert Systems with Applications*, 36: 7758-7763, 2009.
- [11] Coello, C.; A short tutorial on evolutionary multi-objective optimization, *Computer Science*, 1993: 21-40, 2001.
- [12] Fritz Institute; *Lessons learned: recipient perceptions of aid effectiveness: rescue, relief and rehabilitation in tsunami affected Indonesia, India and SriLanka*, San Francisco, CA: Fritz Institute, 2005.
- [13] Goldberg, D.; *Genetic algorithms in search, optimization, and machine learning*, Advison-Wesley, 1989.

-
- [14] Harbaoui, D.I. et al.; Multi-Objective Optimization for the m-PDPTW: Aggregation Method With Use of Genetic Algorithm and Lower Bounds, *Int J Comput Commun*, ISSN 1841-9836, 6(2):246-257, 2001.
- [15] Holland, J.; *Adaptation in natural and artificial systems*, Tech. rep., University of Michigan Press, Ann Arbor, Canberra ACT 2601, Australia, 1975.
- [16] Hong, L.; An improved LNS algorithm for real-time vehicle routing problem with time windows, *Computers and Operations Research*, 39(2):151-163, February, 2012.
- [17] Kefi, G.M.; Ghedira, K.; A Multi-Agent Model for a Vehicle Routing Problem with Time Windows, Urban Transport Conference, Dresden-Allemagne, 2004.
- [18] Kovacs, G.; Spens, K.; Humanitarian logistics in disaster relief operations, *International Journal of Physical Distribution and Logistics Management*, 36(2): 99-114, 2007.
- [19] Long, D.; Wood, D.; The logistics of famine relief, *Journal of Business Logistics*, 16(1): 213-229, 1995.
- [20] Ma, X. et al.; Min-max robust optimization for the wounded transfer problem in large-scale emergencies, Control and Design Conference, China, 2010.
- [21] McClintock, A.; The logistics of humanitarian emergencies: notes from the field, *Journal of Contingencies and Crisis Management*, 17(4): 295-302, 2009.
- [22] McEntire, D.; Issues in disaster relief: progress, perpetual problems and prospective solutions, *Disaster Prevention and Management*, 8(5): 351-361, 1999.
- [23] McGuire, G.; Supply chain management in the context of international humanitarian assistance in complex emergencies, *Supply Chain Practice*, 2(4): 30-43, 2000.
- [24] Mguis, F. et al.; Modlisation dun systme multi-agent pour la rsolution dun problme de tournes de vhicules dans une situation durgence, in: 9me Confrence Internationale de Modlisation, Optimisation et SIMulation MOSIM12, Bordeaux, France, 2012.
- [25] Mguis, F. et al.; Distributed approach for vehicle routing problem in disaster case, 13th IFAC Symposium on Control in Transportation Systems, Sofia-Bulgaria, 2012.
- [26] Mguis, F. et al.; Guided genetic algorithm for the dynamic management of emergency planning for disaster, Internationnal conference of Information Technology and Quantitative Management, Suzhou-China, 2013.
- [27] Oh, S.; Haghani, A.; Testing and evaluation of a multi-commodity multi-modal network flow model for disaster relief management, *Journal of Advanced Transportation*, 31: 249-282, 1997.
- [28] Oloruntoba, R.; Gray, R.; Customer service in emergency relief chains, *International Journal of Physical Distribution and Logistics Management*, 39(6): 486-505, 2009.
- [29] Oloruntoba, R.A.; Documentary analysis of the cyclone Larry emergency relief chain: some key success factors, *International Journal of Production Economics*, 126(1): 85-101, 2010.
- [30] Ozdamar, L.; Planning helicopter logistics in disaster relief, *OR Spectrum*, 33: 655-672, 2011.

-
- [31] Ozdamar, L.; Demir, O.; A hierarchical clustering and routing procedure for large scale disaster relief logistics planning. *Transportation Research Part E: Logistics and Transportation Review*, 48: 591-602, 2012.
- [32] Perry, M. ; Natural disaster management planning: a study of logistics managers responding to the tsunami, *International Journal of Physical Distribution & Logistics Management*, 37(5): 409-433, 2007.
- [33] Petitt, S.; Beresford, A.; Emergency relief logistics: an evaluation of military, nonmilitary and composite response models, *International Journal of Logistics: Research and Applications*, 8: 313-331, 2005.
- [34] Psaraftis, H.N.; *Dynamic vehicle routing problems*, *Vehicle routing: methods and studies*, Elsevier Science Publishers B.V.: 293-318, 1988.
- [35] Psaraftis, H.N.; Dynamic vehicle routing: Status and prospects, *Annals of Operations Research*, 143-164, 1995.
- [36] Quarantelli, E.; Ten research derived principles of disaster planning, *Disaster Management*, 2: 23-26, 1982.
- [37] Quarantelli, E.; *What is a disaster*, London: Routledge, 1998.
- [38] Rosenthal, U. et al.; *Coping with crises: the management of disasters, riots and terrorism*, Springfield, IL: Charles C. Thomas Publishers, 1989.
- [39] Savvaiddis, P.et al); Organization of emergency response after a major disaster event in an urban area with the help of an automatic vehicle location and control system, *GPS Solutions*, 5(4): 70-79, 2002.
- [40] Solomon, M.; Algorithms for the vehicle routing and scheduling problems with time window constraints, *Operations Research*, 35(2): 254-265, 1987.
- [41] Tzeng, G.H.et al.; Multi-objective optimal planning for designing relief delivery systems, *Transportation Research Part E: Logistics and Transportation Review*, 43: 673-686, 2007.
- [42] Yi, W.; Kumar, A. (2007); Ant colony optimization for disaster relief, operations, *Transportation Research Part E: Logistics and Transportation Review*, 43 (6): 660-672, 2007.
- [43] Zeddini, B.; Zargayouna M.; *Auto-organisation spatio-temporelle pour le VRPTW dynamique*, RJCIA, 2009.
- [44] Zhang, J.H.et al.; Multiple-resource and multiple-depot emergency response problem considering secondary disasters, *Expert Systems with Applications* ,39, 11066-11071, 2012.
- [45] Zidi, I. et al.; A Multi-Agent System based on the Multi-Objective Simulated Annealing Algorithm for the Static Dial a Ride Problem, 18th World Congress of the International Federation of Automatic Control (IFAC), Milano (Italy), 2011.
- [46] Zidi, K.; *Système Interactif d'Aide au Déplacement Multimodal*, Thèse de doctorat Ecole centrale de Lille France, 2006.
- [47] <http://www.cbsnews.com/stories/2010/01/13/world/main6090601.shtml>
- [48] <http://www.reliefweb.int>

Author index

An J., 760
Andonie R., 689

Bai J., 654
Behardien S., 708
Bessa W.M., 736
Bocu D., 662
Bocu R., 662
Borne P., 769
Brodić D., 673
Bucerzan D., 681

Caçaron A., 689
Chueh Y., 689

Deng S., 744
Deng Y., 744

El Ferchichi S., 699

Fernandes J.M.M., 736

Ghedira K., 769

Hu F., 760
Huang H., 744

Jing S., 754

Kruger C., 708
Ksouri M., 699

Laabidi K., 699
Liu W., 754

Maluckov Č.A., 673
Manolescu M.J., 681
Maouche S., 699
Mguis F., 769

Nang J., 722

Oh B., 722

Park S., 722

Peng L., 673

Ratiu C., 681
Retonda-Modiya J., 708

Tanaka M.C., 736

Wang F., 754
Wu P., 754
Wu Y., 744

Yang J., 722
Yao J., 760
Yu J., 722

Zidi K., 769
Zidi S., 699