Lecture Notes in Electrical Engineering 247

Haeng Kon Kim
Sio-Iong Ao
Mahyar A. Amouzegar
Burghard B. Rieger
Editors

IAENG Transactions on Engineering Technologies

Special Issue of the World Congress on Engineering and Computer Science 2012

Springer

Lecture Notes in Electrical Engineering

Volume 247

For further volumes:
http://www.springer.com/series/7818

Haeng Kon Kim · Sio-Iong Ao
Mahyar A. Amouzegar
Burghard B. Rieger
Editors

IAENG Transactions on Engineering Technologies

Special Issue of the World Congress on Engineering and Computer Science 2012

Springer

Editors
Haeng Kon Kim
Engineering College, Department of Computer and Communication
Catholic University of DaeGu
DaeGu
Korea, Republic of South Korea

Sio-Iong Ao
International Association of Engineers
Hong Kong
Hong Kong SAR

Mahyar A. Amouzegar
College of Engineering
California State Polytechnic University
Pomona, CA
USA

Burghard B. Rieger
Inst. Computerlinguistik, Abt. Linguistische Datenverarbeitung
Universität Trier
Trier
Germany

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-94-007-6817-8 ISBN 978-94-007-6818-5 (eBook)
DOI 10.1007/978-94-007-6818-5
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013945283

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in San Francisco, California, USA, October 24–26, 2012, under the World Congress on Engineering and Computer Science (WCECS 2012). The WCECS 2012 is organized by the International Association of Engineers (IAENG). IAENG is a non-profit international association for engineers and computer scientists, which was founded originally in 1968 and has been undergoing rapid expansions in the recent years. The WCECS congress serves as an excellent platform for the engineering community to meet with each other and to exchange ideas. The congress has also struck a balance between theoretical and application development. The conference committees have been formed with over 200 members who are mainly research center heads, deans, department heads/chairs, professors, and research scientists from over 30 countries. The full committee list is available at the congress' web site: www.iaeng.org/WCECS2012/committee.html. The congress is truly an international meeting with a high level of participation from many countries. The response to the conference call for papers was excellent with more than 800 manuscript submissions for WCECS 2012. All submitted papers went through the peer review process and the overall acceptance rate was 53.16 %.

This volume contains 49 revised and extended research articles, written by prominent researchers participating in the conference. Topics covered include circuits, engineering mathematics, control theory, communications systems, systems engineering, manufacture engineering, computational biology, chemical engineering and industrial applications. This book offers the state of art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent source of reference for researchers and graduate students working with/on engineering technologies and physical science and applications.

Haeng Kon Kim
Sio-Iong Ao
Mahyar A. Amouzegar
Burghard B. Rieger

Contents

Chapter 1
A 480 MHz Band-Pass Sigma Delta Analog to Digital Modulator with Active Inductor Based Resonators

Kevin Dobson, Shahrokh Ahmadi and Mona Zaghloul

Abstract This chapter presents a 480 MHz, continuous time, 6th order band-pass Sigma Delta Analog to Digital modulator in IBM 0.18 um CMOS technology. We replace traditional RLC circuits, containing spiral inductors with high quality factor, active inductor based resonators utilizing negative impedance circuits. This reduces chip area and eliminates post processing needs. Pad to pad simulation of the extracted layout in Cadence yields an enhanced SNDR of 70 dB and power consumption of 29 mW. The modulator occupies 0.5 mm^2 of chip area.

Keywords Active inductor · Analog to digital converter · Negative impedance circuit · Sigma delta · Sixth order

1.1 Introduction

Analog to Digital Converters (ADCs) allow us to convert analog signals to digital representations suitable for processing by a digital computer. Sigma-Delta ($\Sigma\Delta$) modulators utilize the processes of oversampling and noise shaping in order to obtain high Signal to Noise Ratios (SNR). $\Sigma\Delta$ modulators utilize a few critical components and produce high accuracy results [1]. These are highly desirable

K. Dobson (✉) · S. Ahmadi · M. Zaghloul
Department of Electrical and Computer Engineering, George Washington University,
801 22nd Street NW, Washington DC 20052, USA
e-mail: kdobson@gwu.edu

S. Ahmadi
e-mail: ahmadi@gwu.edu

M. Zaghloul
e-mail: zaghloul@gwu.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_1,

attributes. The relationship between the SNR measured at the output of a ΣΔ modulator and the effective number of bits (ENOB) is given by:

$$SNR = 6.02ENOB + 1.76 \tag{1.1}$$

Continuous Time (CT) ΣΔ modulators offer inherent antialiasing and are able to operate at higher frequencies than their discrete time counterparts.

In Radio Frequency (RF) receivers an incoming signal is repeatedly filtered and mixed down to lower frequencies before being digitized and processed. A CT band-pass ΣΔ modulator capable of digitizing an RF carrier signal would eliminate the need for analog filtering and mixing, and these functions would be passed on to a Digital Signal Processor (DSP). This results in a simpler, cheaper and more efficient receiver [2].

1.2 Design of Continuous Time Sigma Delta Modulator

The general architecture of a CT ΣΔ modulator is shown in Fig. 1.1. When designing the loop filter $G(s)$ for a CT ΣΔ modulator we begin with a discrete time modulator transfer function $F(z)$.

Once $F(z)$ has been chosen the impulse invariant method can be used to determine the equivalent CT loop filter $G(s)$ as outlined in Lelandais-Perrault et al. [3]. For modulators with a non-return-to-zero (NRZ) feedback we get:

$$F(z) = (1 - z^{-1})Z_T\{L^{-1}[\frac{G(s)e^{-ds}}{s}]\} \tag{1.2}$$

Here d represents the delay introduced by the ADC and Digital to Analog Converter (DAC). For a sixth order band-pass Sigma Delta modulator the equivalent continuous time loop filter transfer function is of the form:

$$G(s) = \frac{(s-a)(s^2 + \frac{\omega_a}{Qa}s + \omega_a^2)(s^2 + \frac{\omega_b}{Qb}s + \omega_b^2)}{(s^2 + \frac{\omega_0}{Q_0}s + \omega_0^2)(s^2 + \frac{\omega_1}{Q_1}s + \omega_1^2)(s^2 + \frac{\omega_2}{Q_2}s + \omega_2^2)} \tag{1.3}$$

ω is the normalized resonator frequency with respect to the sampling frequency in radians per second, and Q is the quality factor of the resonators. The sampling frequency is T. When d is equal to 1.4 T and the sample rate is 4 times the frequency of the input signal then the term a in the numerator approximates to zero [4].

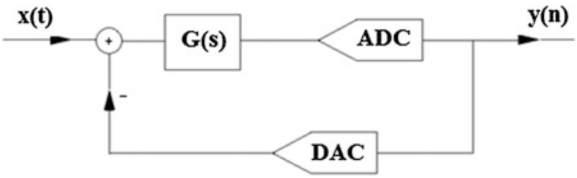


Fig. 1.1 Continuous time sigma delta modulator

1.3 Active Inductor Resonator Structure

A parallel RLC resonator is shown in Fig. 1.2.

The transfer function of a parallel RLC circuit resonator $H(s)$ is given by:

$$H(s) = \frac{As}{s^2 + \frac{\omega_0}{Q} + \omega_0^2}$$

$$\text{where } \omega_0 = \frac{1}{\sqrt{LC}} \text{ and } Q = R_P C\omega_0 = \frac{R_P}{L\omega_0} \tag{1.4}$$

$G(s)$ cannot be realized as a cascade of resonators but can be realized by the structure in Fig. 1.3. Here g, A_H and A_L represent amplifier gains, and H the resonators. The Σ block is an analog adder.

Traditionally resonators for band-pass CT ΣΔ modulators have been realized by RLC parallel circuits with spiral inductors. Such circuits occupy a large silicon area. Spiral inductors also have low quality factors. Active inductor based RLC circuits occupy a much smaller area, and when Q enhancement techniques are used, high quality factors can be achieved.

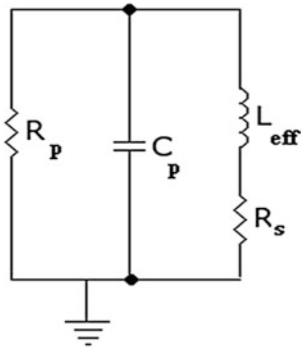


Fig. 1.2 RLC resonator

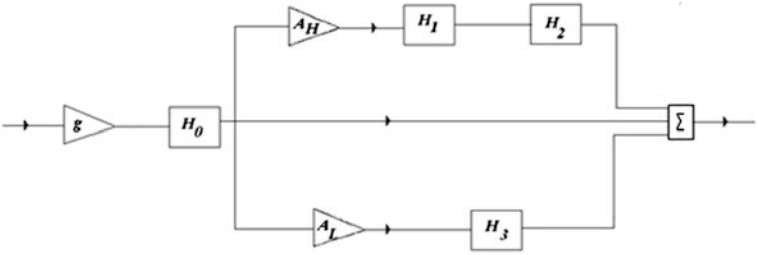


Fig. 1.3 Sixth order active inductor based loop filter

The active inductor based resonator is explained by the gyrator C theorem as shown in Fig. 1.4.

$$I' = -V_{in} G_{m1} \tag{1.5}$$

$$-I_{in} = V' G_{m2} \tag{1.6}$$

$$V' = I' \frac{1}{sC} \tag{1.7}$$

After substitution we get:

$$\frac{V_{in}}{I_{in}} = \frac{sC}{G_{m1} G_{m2}} \tag{1.8}$$

In Eq. (1.8) we note that the s is in the numerator indicating that the circuit is inductive. G_{m1} and G_{m2} can be realized using CMOS devices.

The circuit in Fig. 1.5 realizes an active inductor with M1 and M2 acting as G_{m1} and G_{m2} respectively.

A detailed small signal analysis results in an expression for Z_{in} as shown in Eq. (1.9) below [5].

Fig. 1.4 Gyrator topology

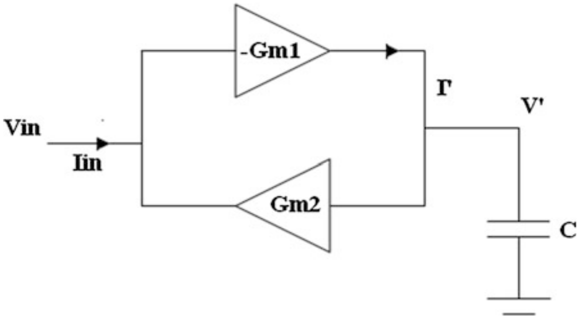
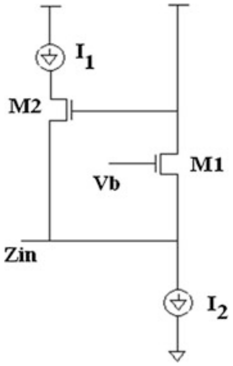


Fig. 1.5 CMOS active inductor



$$Z_{in} = \frac{g_{oc} + g_{o1} + s(C_{gs2} + C_{gd2} + C_{ds1})}{g_{m1}g_{m2} + [g_{m2} - g_{m1} + g_{oc} + s(C_{gs2} + C_{ds1})](g_{o2} + sC_{gd2})} \tag{1.9}$$

Here g_o is the drain-source conductance and g_{oc} represents the loading effect of the non-ideal biasing current source. Z_{in} can be interpreted to represent the parallel RLC circuit as shown in Fig. 1.2.

Separating the Resistive, Capacitive and inductive parts of Eq. (1.9) yields the following:

$$R_p = \frac{1}{g_{m1}} \tag{1.10}$$

$$C_p = C_{gs1} \tag{1.11}$$

$$L_p = \frac{C_{gs2}}{g_{m1}g_{m2}} \tag{1.12}$$

$$R_s = \frac{g_{oc} + g_{o1}}{g_{m1}g_{m2}} \tag{1.13}$$

The intrinsic self-resonant frequency and intrinsic quality factor of the circuit is given respectively by:

$$\omega_0 = \sqrt{\frac{g_{m1}g_{m2}}{C_{gs1}C_{gs2}}} \tag{1.14}$$

$$Q_0 = \sqrt{\frac{g_{m2}C_{gs1}}{g_{m1}C_{gs2}}} \tag{1.15}$$

By utilizing two similar circuits to that in Fig. 1.5 and a Negative Impedance Circuit (NIC), a high Q fully differential resonator can be designed for use in a band-pass ΣΔ modulator. This resonator is shown in Fig. 1.6. Output buffers are used but not shown.

A PMOS device is used to couple the input to the circuit. It draws a small amount of current and does not disturb the gyrator function. Output gain can be controlled by varying the size of this MOSFET. The effect of cascoding M3 with M2 reduces the output conductance thereby reducing R_s and increasing the Q. The NIC is comprised of 3 cross-coupled differential pairs of MOSFETs with drains tied to the opposing gates. It provides a negative resistance that seeks to cancel the parallel resistance R_p, further increasing the Q. When one cross coupled pair of MOSFETs is used as a NIC, it provides a negative resistance of $-2/\mathrm{g_m}$ and adds a $\mathrm{C_{gs}}/2$ parasitic shunt capacitance [6]. These simple NICs however, are notoriously nonlinear. In order to obtain greater linearity a multi-tanh version of the NIC circuit was used. This requires the addition of two extra cross coupled pairs of MOSFETs with a 2:1 size ratio [7, 8]. When the signal is large and the

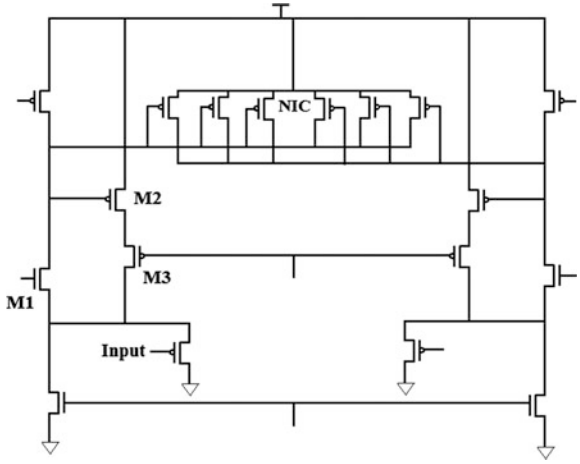


Fig. 1.6 CMOS active inductor based resonator with NIC

symmetrical differential pair has saturated, the unbalanced differential pairs can still provide a differential current proportional to the input voltage. This scheme works effectively at high frequencies. We can rewrite Q_0 as:

$$Q_0 = \frac{1}{g_{m1}} \sqrt{\frac{C_p}{L_p}} \tag{1.16}$$

If we denote the enhanced quality factor of the circuit with a NIC asQ_n, then;

$$Q_n = \frac{1}{g_{m1} - g_{nic}} \sqrt{\frac{C_p + C_{nic}}{L_p}} \tag{1.17}$$

Here $\mathrm{g_{nic}}$ is the NIC transconductance and $\mathrm{C_{nic}}$ is the capacitance the NIC adds to the circuit [7]. As can be seen from Eq. (1.17) the closer the transconductance of M1 and the NIC the higher the Q. Care must be taken during design to ensure that the NIC transconductance does not exceed the transconductance of M1. The added NIC capacitance decreases the resonant frequency. This can be compensated for by increasing biasing currents.

While there is no limit to the voltage that can be applied to spiral inductors, the maximum input voltage to active inductor based circuits must not cause MOSFETs to cease operating in saturation mode. Active inductor based circuits are also noisier than circuits with real inductors by a factor of $2Q_0$ [7].

We were able to design and simulate the schematics of active inductor based resonators in Cadence with resonant frequencies between 100 MHz and 1.5 GHz and were consistently able to achieve a Q of 50 or greater.

1.4 Simulation of Continuous Time Sigma Delta Modulator

A Matlab program was used to generate initial values of the resonator multiplying coefficients g, A_H and A_L that fulfilled $G(s)$. Pole-Zero plots were then done to confirm modulator stability. Next, Simulink simulations were used to further refine the modulator design. The Simulink model was easily modified to reflect the non-idealities of an actual circuit such as limited gain due to nonlinearity and small delays introduced by each circuit component. In the ideal case with high gain in the path containing the most resonators a theoretical Signal to Noise-plus-Distortion Ratio (SNDR) of 95 dB was obtained. When the limitations of nonlinearity and circuit delay were considered a goal of 75 dB for operation at 1.2 GHz was settled on. This required a 300 MHz resonator. For this schematic design, linear operation was observed when the input was limited to less than 10 mv p-p. The Q was 50.

Cadence simulation of the designed schematic followed. A block diagram of the complete circuit simulated in Cadence is shown in Fig. 1.7.

A series of comparators [9] A, are used to provide the required delay of 1.4T and effective amplification of the signal prior to quantization. A large enough signal at the input of the quantizer is necessary to prevent clock feed-through [9]. Since there is a non-zero delay it is necessary to add a direct loop between the DAC and the ADC input [1]. The Adder used is described in [4].

A Differential pair is used to subtract the DAC output from the input signal as shown in Fig. 1.8.

A clocked comparator [9] coupled with a SR flip flop is used to generate the modulator output. The flip flop is necessary because a NRZ output is required.

The DAC shown in Fig. 1.9 converts the rail to rail output swing of the quantizer to a smaller voltage equal to the maximum peak to peak analog input.

Cadence schematic simulation results were similar to Simulink simulations in Fig. 1.10, albeit with deteriorated noise shaping due to the circuit non-idealities previously mentioned and others such as settling time of the adder output, and offset errors. Nevertheless both yielded a SNDR of 75 dB as outlined in [10].

The layout of the schematic previously simulated was created and extracted. Upon simulation of the extracted layout it was evident that there needed to be some circuit modifications in order to achieve a working modulator. During the schematic simulation, the impedance of interconnecting wires was ignored.

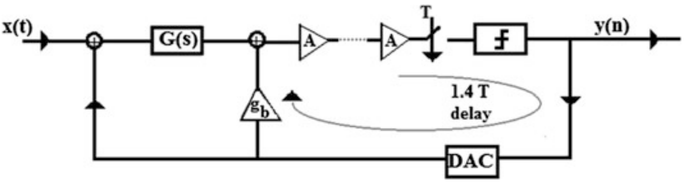


Fig. 1.7 Active inductor based sixth order continuous time modulator

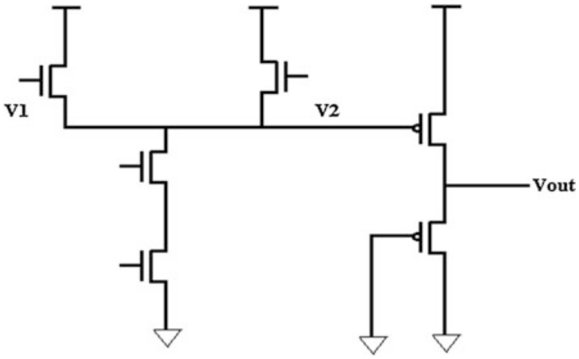


Fig. 1.8 Subtracter differential pair with buffered output

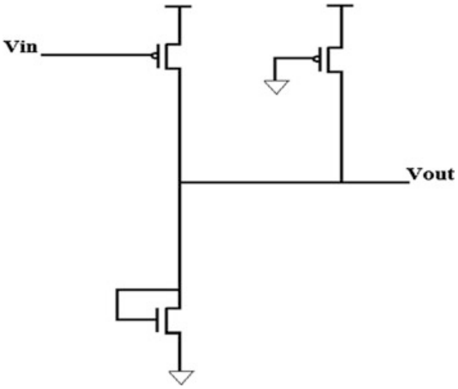


Fig. 1.9 DAC

Simulation of the extracted view takes these impedances into account. The interconnecting wire leading to the active inductor and the gate capacitances of the input PMOS device form a voltage divider. In order to couple input signals effectively to the active inductor, while maintaining the required gains previously determined, it is necessary to make the PMOS devices larger so that the capacitances seen at the gates of the PMOS devices are large and will effectively couple most of the input voltage to the active inductor.

The addition of parasitic series resistances and parallel capacitances to the circuit results in a lower resonant frequency and decreased Q. The extra parasitics also result in greater propagation delays between sub-circuits. This limits the modulator clock speed.

By increasing the size of M3 we decrease R_s which increases our deteriorated Q. Since G_{m1} and G_{m2} form a feedback loop, instability occurs when G_{m2} gets too high i.e. when M3 is too large [11].

The output buffers between active inductors have to be slightly adjusted in order to maintain bias points.

Because we have both digital and analog circuits on the same substrate we use separate VDD and GND for these sub-circuits. We also surround our analog

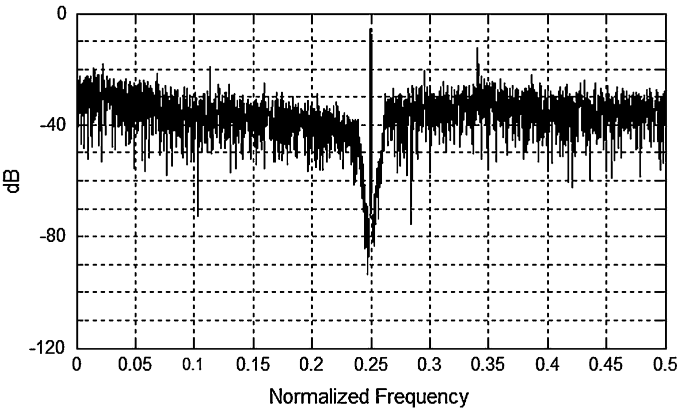


Fig. 1.10 Simulink modulator output power spectrum density

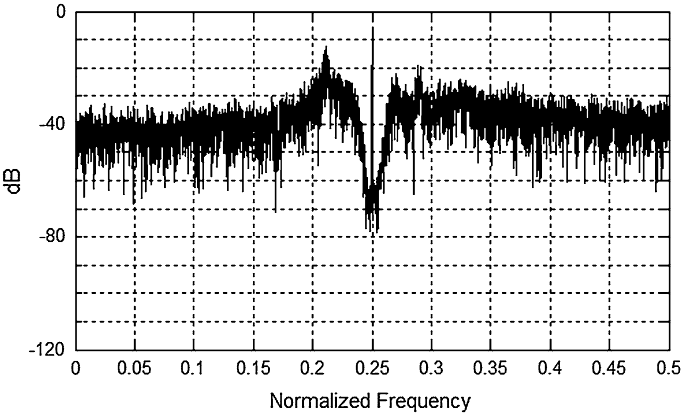


Fig. 1.11 Modulator output power spectrum density from Cadence pad to pad extracted layout simulation

sub-circuits with two guard rings separated by BFMOAT in order to protect them from digital noise [12].

The changes discussed above were made to the layout and Input/Output pads added. Figure 1.12 shows this layout. The layout was re-extracted and a pad to pad simulation was conducted. We were able to achieve a SNDR of 70 dB for a modulator clocked at 480 MHz with 120 MHz resonators. This result is shown in Fig. 1.11.

This still compares favorably with other non-active inductor based sixth order band-pass ΣΔ modulators such as [13], which yields a SNDR of 68 dB. Our circuit consumes 29 mW which is much smaller than the 160 mW consumed in [13] and occupies a mere 0.5 mm^2, as compared to the 2.5 mm^2 used consumed in [13].

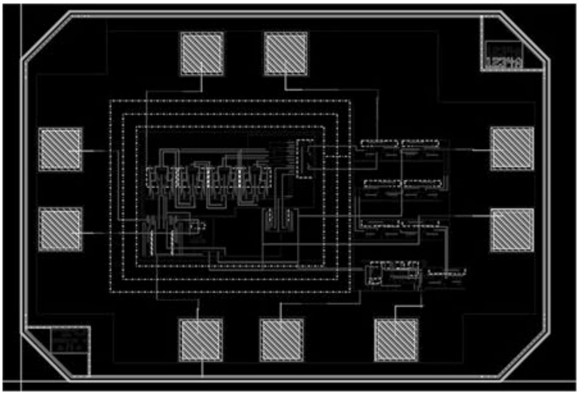


Fig. 1.12 Layout of 480 MHz band-pass sigma delta analog to digital modulator with active inductor based resonators

1.5 Conclusion

We have succeeded in designing and simulating the first Sixth Order, CT ΣΔ modulator using active inductor based resonators in the loop filter. The use of Q enhancing techniques has resulted in a modulator with a high SNDR, and we have avoided the use of area consuming spiral inductors. When compared to the 47 dB, fourth order, active inductor based CT ΣΔ mentioned in [14] we are able to achieve a greater SNDR and consume roughly the same amount of power. Our design goes beyond schematic simulation and tackles and overcomes the real life design issues encountered when laying out this novel architecture.

References

1. Benabes P, Keramat M, Kielbasa R (1998) Synthesis and analysis of sigma-delta modulators employing continuous-time filters. Analog Integr Circ Sig Process 23:141–152
2. Schreier R, Temes G (2005) Understanding delta-sigma data converters. IEEE press, Piscataway
3. Lelandais-Perrault C, Benabes P, De Gouy J, Kielbasa R (2003) A parallel structure of a continuous-time filter for bandpass sigma-delta A/D Converters. In: Proceedings of 10th IEEE international conference on electronics, Sharjah (Emirates Arabes Unis)
4. Benabid S, Benabes P (2003) High linear integrated LC filter for a continuous-time bandpass sigma-delta ADC circuits and systems, vol 1(30). 2003 IEEE 46th Midwest symposium, pp 291–294, Dec 2003
5. Gao Z, Yu M, Ye Y, Ma J (2006) A CMOS bandpass filter with wide-tuning range for wireless applications, circuits and systems. ISCAS 2006. In: Proceedings of 2006 IEEE international symposium
6. Jung B, Harjani R (2004) A wide tuning range VCO using capacitive source degeneration, circuits and systems. ISCAS '04. In: Proceedings of the 2004 international symposium, vol 4, pp IV–145-8, 23–26 May 2004
7. Wu Y, Ding X, Ismail M, Olsson H (2003) RF bandpass filter design based on CMOS active inductors, IEEE transactions on circuits and systems—II: analog and digital signal processing, vol 50, no. 12, Dec 2003

8. Ryan AP, McCarthy O (2004) A novel pole-zero compensation scheme using unbalanced differential pairs, IEEE transactions on circuits and systems—I: regular papers, vol. 51, no. 2, Feb 2004
9. Baker RJ (2011) CMOS, Circuit design, layout, and simulation, IEEE press series on microelectronic systems. Wiley, New Jersey
10. Dobson K, Ahmadi S, Zaghloul M (2012) A 1.2 GHz band-pass sigma delta analog to digital modulator with active inductor based resonators. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, USA, pp 875–879, 24–26 Oct 2012
11. Bakken T, Choma J (2002) Gyrator-based synthesis of active on-chip inductances. J Analog Integr Circ Sig Process 34:171–181
12. Dai H (2008) Differential sensing of substrate noise in mixed-signal 0.18-um BiCMOS technology. Electron Device Lett IEEE 29(8):898–901
13. Lu C-Y, Silva-Rivas JF, Kode P, Silva-Martinez J, Hoyos S (2010) A sixth-order 200 MHz IF bandpass sigma-delta modulator with over 68 dB SNDR in 10 MHz bandwidth. Solid-State Circ IEEE J 45(6):1122–1136
14. Chen Q, Sankary KEl, Masry EEl (2008) A UHF continuous-time current-mode band-pass delta sigma modulator based on active inductor, circuits and systems. MWSCAS 2008. In: 51st Midwest symposium

Chapter 2
Basic Computations Using a Novel Scalable Pulse-Mode Modules

Thamira Hindo

Abstract In this chapter the basic computational functions used in many algorithms are implemented in pulse mode. For this purpose, a novel circuit is proposed for pulse-based logarithmic computation using integrate-and-fire (IF) structures. The smallest unit in the module is a network of three IF units that implements a margin propagation (MP) function using integration and threshold operations inherited in the response of an IF neuron. The three units are connected together through excitatory and inhibitory inputs to impose constraints on the network firing-rate. The MP function is based on the log likelihood computation in which the multiplication of the inputs is translated into a simple addition. The advantage of using integrate-and-fire margin propagation (IFMP) is to implement a complex non-linear and dynamic programming functions of spike based (pulse based) computation in a modular and scalable way. In addition to scalability, the objective of the proposed module is to map algorithms into low power circuits as an attempt to implement signal processing applications on silicon. The chapter shows the mechanism of IFMP circuit, dynamic characteristics, the cascaded modularity, the verification of the algorithm in analog circuit using standard 0.5μm CMOS technology and the basic functions computation.

Keywords Excitatory · Inhibitory · Integrate and fire · log-sum-exp · Margin propagation · Pulse mode computation.

T. Hindo (✉)
Department of Electrical and Computer Engineering, Michigan State University,
East Lansing, MI 48824, USA
e-mail: hindotha@msu.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_2,

2.1 Introduction

Although Von Neumann computer architecture perform high speed computation and communication but they are unable to perform brain tasks processes in an efficient way such as the biological sensing in the retina and cochlea. A new trend in computer architecture for applications other than precise, high speed calculation and efficient communications is now in the fourth generation of research work to built up a neuromorphic systems. The goal of the neuromorphic systems is to implement sensory devices in an efficient way as in the biological sensors [1–5]. The architecture of the morphed biological systems are different from the traditional Von Neumann architecture such as asynchronous- parallel processing instead of synchronous- single processing, hybrid computation instead of digital computation, neuron model as a basic core of the processing instead of the arithmetic logic unit and finally, analog VLSI design instead of digital VLSI. As a contribution in this huge project, a novel and scalable algorithm is proposed to approximate non-linear function as an important procedure to implement signal processing algorithms in the sensory applications such as recognition and classification. The main objective of this work is to map a pulsed mode algorithm into low power silicon circuits as an attempt to implement signal applications in the neuromorphic systems.

The proposed module has three concepts, the first concept is to map the non-linear functions into margin propagation “MP” [6] which is an approximation function to the log-sum-exp (LSE) expression. The second concept of the proposed module is based on an integrate and fire neuron model. The third concept is to implement the non-linear function in pulse stream mode. The proposed pulse mode computation module is abbreviated as “IFMP” since it implements a MP function using IF neuron model.

For the first concept, the LSE math function is used in factor graph algorithms in which the sum of product terms are used [7]. In these algorithms, the probabilities or the marginal functions of passing messages are evaluated between the nodes and variables of a factor graph. Since the product of probability terms tend to decrease as the number of probability terms increase, then we would have a problem of underflow that cause false computations. Therefore, such algorithms use the log-likelihood computation to eliminate the underflow problem as well as to increase the dynamic range of variables in the computation process. But the representation of LSE is not scalable in hardware design. Therefore the margin propagation function (scalable function in hard ware design) is used as an approximation method to the LSE function. The concept of margin propagation (MP) algorithm is based on the idea of reverse water-filling (RWF) algorithm, [8]. Given a set of random inputs (scores) $L_i \in R; i = 1 : m$, the RWF algorithm computes the solution z according to the constraint,

$$\sum_{i=1}^{m} [L_i - z]_+ = \gamma \tag{2.1}$$

where $[.]_+ = max(.,0)$ denotes a threshold operation and $\gamma \geq 0$ represents a parameter of the algorithm. Note that z is in the log domain. The solution of the equation 2.1 is represented by z, where z can be written in LSE and MP forms as,

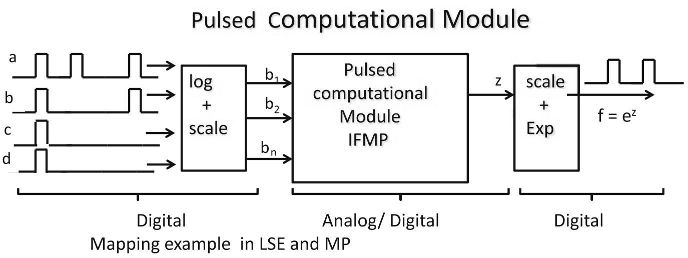
$$z = log(\sum_{i=1}^{m} e^{L_i} \simeq M(L_1, L_2, L_3 \ldots L_m, \gamma) \tag{2.2}$$

where M denotes as the MP function, m denotes the number of the input operand. In previous work [9], it was proven that MP is successfully an approximation method to LSE. The input/output variables in the above work are represented as currents, also the computation procedure is implemented using kirchoffs current law. MP was implemented in [9] to achieve scalability in the decoding algorithms. In this work, we introduce the concept of MP propagation as an approximation method to LSE in pulse computation mechanism.

Secondly, the structure of the proposed module is based on an integrate and fire (IF) neuron model that implement integration and threshold operation. Since the IF model is a simple representation of a neuron, it is extensively used as a neuron model in spiking neural networks [10–12] and neuromorphic systems [4, 13]. The IF neuron itself is the basic computational unit in the sophisticated and efficient architecture, "the brain".

The brain is the most realistic example of an efficient system, "hybrid system", which is the third concept of the proposed module. The type of the signals (data) transferred in the brain is mixed between digital as spikes (pulses) and analog as the variable time between these spikes. The above signal processing is called pulse stream mode or hybrid computations [14, 15]. Hybrid (pulse) computation is a promising research topic since it mixes the advantages of analog and digital designs. The noise accumulation in analog stages can be eliminated by digital noise immunity. The analog design has the advantage of small area, low cost and low power especially if the design of computational units is implemented in weak inversion mode of complementary metal oxide semiconductors (CMOS). Figure 2.1 shows the flow of the proposed module which includes manipulating the input pulse rate (scaling and converting into logarithm domain), mapping the function into MP and evaluate the output z, then scaling back and calculating the exponent of z to realize the function.

The concept of the proposed algorithm is analyzed, mapped and verified into a low power analog circuit, verify the properties of IFMP and the computations of the basic functions that are the core of signal processing algorithms and finally verified the concept of pulse computation on 0.5μm process chip. This chapter is organized as following: Sect. 2.2 demonstrates the analysis, synthesis, dynamic characteristics of IFMP module. Section 2.3 explains the circuit description and the hardware verification of IFMP. Section 2.4 shows the chip test implementation. Section 2.5 describes the properties of the IFMP modules and how they are used to implement computational functions. Section 2.6 concludes the chapter with the future work.



$$f = ab + cd$$

$$\log(f) = \log(ab + cd)$$

$$\log(f) = \log(e^{\log(ab)} + e^{\log(cd)})$$

$$LSE: z = \log[e^{\log(a)+\log(a)} + e^{\log(c)+\log(d)}]$$

$$MP: z = M(\log(a) + \log(b), \log(c) + \log(d), \gamma)$$

Fig. 2.1 Input/output stages in the pulsed computational module

2.2 IFMP: Concepts and Analysis

Figure 2.2 shows a schematic diagram of the proposed spiking network module. The network is referred to as an IFMP module and consists of three integrate-and-fire structures N_1, N_2 and N_3. The excitatory/inhibitory inputs are represented by black/white triangles. Units N_1 and N_2 have self-inhibitory feedback connections and unit N_3 has inhibitory input denoted as γ. Given the rate of input spike-trains $L_1[n]$ and $L_2[n]$ with n being a discrete time-index, it can be shown that firing-rate of the output $L_z[n]$ (denoted by $\mathscr{E}(L_z) = \lim_{T\to\infty} \frac{1}{T}\sum_{n=1}^{T} L_z[n]$) asymptotically satisfies the following equation,

$$[\mathscr{E}(L_1[n]) - \mathscr{E}(L_z[n])]_+ + [\mathscr{E}(L_2[n]) - \mathscr{E}(L_z[n])]_+ \longrightarrow \mathscr{E}(\gamma[n]) \qquad (2.3)$$

and in general for m inputs,

$$\sum_{i=1}^{m} [\mathscr{E}(L_i[n]) - \mathscr{E}(L_z[n])]_+ \longrightarrow \mathscr{E}(\gamma[n]) \qquad (2.4)$$

Note that equation 2.4 converges only in probability. The difference between the left and right hand side of the above equation decreases as the time increases (or the number of stream sequence of random inputs increases) and hence the summation of the expected values of the input stream converges to the expected values of the output stream. In Hindo [16], we proved the convergence of one neuron in order to prove the convergence of IFMP Eq. 2.4. For one neuron, the expected value of output spikes $d[n]$ is equal to the expected value of the input spike $L[n]$ overall the samples as following,

$$\mathscr{E}_n\{L[n]\}_+ = \mathscr{E}_n\{d[n]\} \qquad (2.5)$$

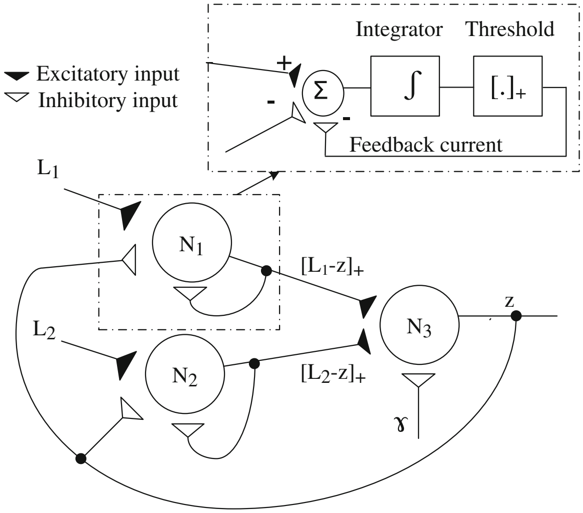


Fig. 2.2 Schematic of the proposed IFMP unit comprising of three integrate-and-fire modules

Figure 2.3 shows the plot of instantaneous spiking-rates for N_1, N_2 and N_3, when the rate of the inputs are varied as shown in Fig. 2.3. In this experiment, $\gamma = 0.3$ and the input rate $L_2 = 0.5$ for N_2 while input rate L_1 for N_1 increases from 0 to 1. The dynamic of the figure follows the IFMP equation 2.4 such that $[L_1 - z]_+ + [L_2 - z]_+ = \gamma$. Initially, when L_1 is between 0 and 0.25, then the output rate of N_1, N_2 and N_3 is equal to 0, 0.3 and 0.2 respectively. When L_1 is 0.3, then the output rate of units N_1, N_2 and N_3 are 0.05, 0.25 and 0.25 respectively. When L_1 is 0.6, then the output rate of units N_1, N_2 and N_3 are 0.2, 0.1 and 0.4 respectively and so on. Hence, the sum of the output rates for the first two IF units N_1, N_2 converges to a constrain rate γ over enough and sufficient time for convergence in which the dynamics of IFMP satisfies equation 2.4 as shown in Fig. 2.3.

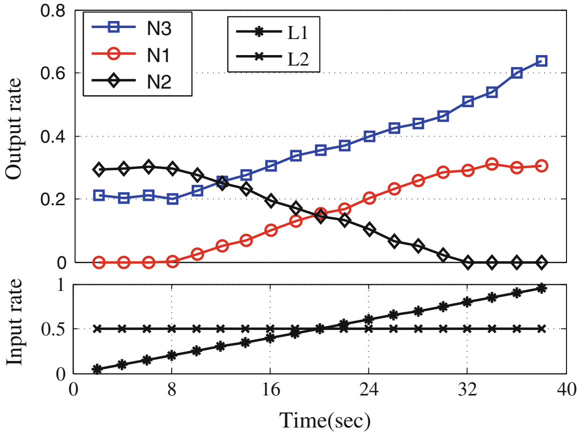


Fig. 2.3 Spike-rates for neurons N_1, N_2 and N_3 (*upper figure*) when the spike-rate of L_1 is monotonically increased, The rate L_2 is kept constant at 0.5 (*lower fig*)

Fig. 2.4 Dynamic characteristics of IFMP unit for different values of γ for MP and IFMP (the analog mode and pulse mode of MP respectively)

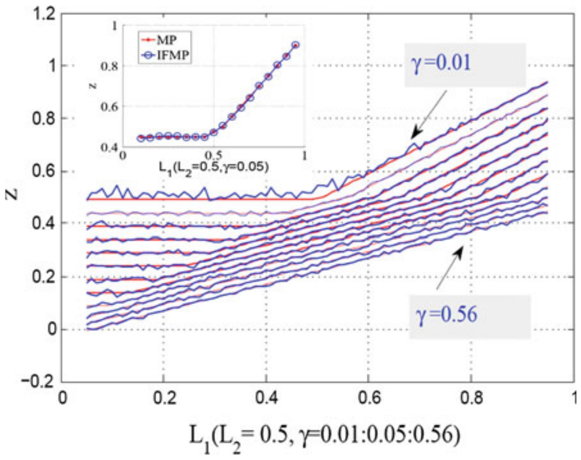


Figure 2.4 shows the plot of instantaneous spiking-rates for N_3, when spiking-rate of the input rate L_1 is varied from 0.01 to 0.9. For this result, the spiking-rate for input L_2 is kept constant to 0.5 as γ changes from 0.01 : 0.05 : 0.56. The plot shows that the spiking-rate of N_3 increases according to a piece-wise linear approximation to the margin propagation function. It was shown in [9] that the margin propagation (MP) is an approximation to the log-sum-exp. However, they did not provide close form representation for the approximation and the parameters involved. Furthermore, they did not demonstrate the efficacy of cascading the approximated model. In Hindo [16], we addressed the parameter involved in the approximation such that $z_{MP} = z_{LSE} - \gamma$ where z_{MP} is the approximated margin propagation value to the LSE value (z_{LSE}). Figure 2.5a shows the approximation which is equal to γ between z_{LSE} and z_{MP}, z_{IFMP} where $z_{IFMP} = z_{MP}$. Furthermore, we showed that the MP formulation can be mapped onto a cascaded topology such that the convergence equation for second IFMP in Fig. 2.5b expressing $L_2 = -\infty$ and derived in Hondo [16] as $[z_{LSE1} - z_{LSE2} + \gamma]_+ + [L_3 - z_{LSE2} + \gamma]_+ = \gamma$, and in general,

$$[z_{LSE[K]} - z_{LSE[K+1]} + \gamma]_+ + [L_{[K+2]} - z_{LSE[K+1]} + \gamma]_+ = \gamma \tag{2.6}$$

Equation 2.6 shows that three inputs MP can be implemented using two identical units of MP which is applicable for higher number of inputs too as shown in Fig. 2.5b. The advantage of cascading is that the algorithms can be implemented using array of 2-IFMP units integrated on silicon while the connectivity could potentially be achieved using a field programmable gate array (FPGA). Therefore, we do not have to redesign the hardware for different applications.

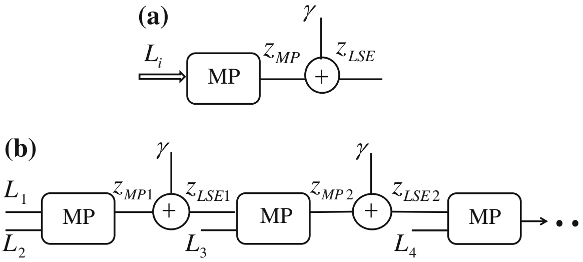


Fig. 2.5 **a** Approximation of IFMP to the LSE math function. **b** Serial cascading or (modularity) of IFPM structure

2.3 IFMP: Circuit Description

The analog circuit of IFMP is designed and shown in Fig. 2.6a, where the shaded area round blocks B1, B2 ,and B3 represent units N1, N2 and N3 of Fig. 2.2 respectively. Block B2 in the upper right of Fig. 2.6a represents the circuit of the integrator and inverter for the three blocks B1, B2 and B3. Figure 2.6b shows the response of the membrane voltage and the convergence equation between the input and output rates for one neuron (represented by block B1 in Fig. 2.6a). Figure 2.6c shows the output pulses (d), membrane voltage (v) and input voltage (in) of the integrator in Block B1.

The integration and threshold are designed between two bounds $(2.34v, 0.9v)$. Initially, if the input of the integrator is zero, the outputs of the integrator and the cascoded inverter are equal to 3.3 and zero volts respectively. If the input voltage increases and reaches the high gain region of integrator amplifier (60 db), then the integration phase will be built which is the discharging phase of the capacitor. The input current is integrated and the output voltage of the integrator discharges to the lower bound. At this point, the output of the cascoded inverter turned into logic one which will turn the output voltage of the integrator to the upper bound (charging phase of capacitor). The cycle of charging and discharging the capacitor C is repeated according to the amount of the current injected to the inputs of the integrator (‘in’ node). The injected currents to the three integrators are applied respectively during off and on states of the input pulses for the excitatory path (PMOS transistors) and inhibitory path (NMOS transistors). Modules N_1, N_2 have two excitatory inputs (PMOS path), one self feedback inhibitory input (NMOS path) and one feedback inhibitory input (NMOS path) from the output of unit N_3, whereas module N_3 has two excitatory inputs (PMOS paths) and one inhibitory input. The last inhibitory input is represented by an adjustable constrain rate γ explained earlier.

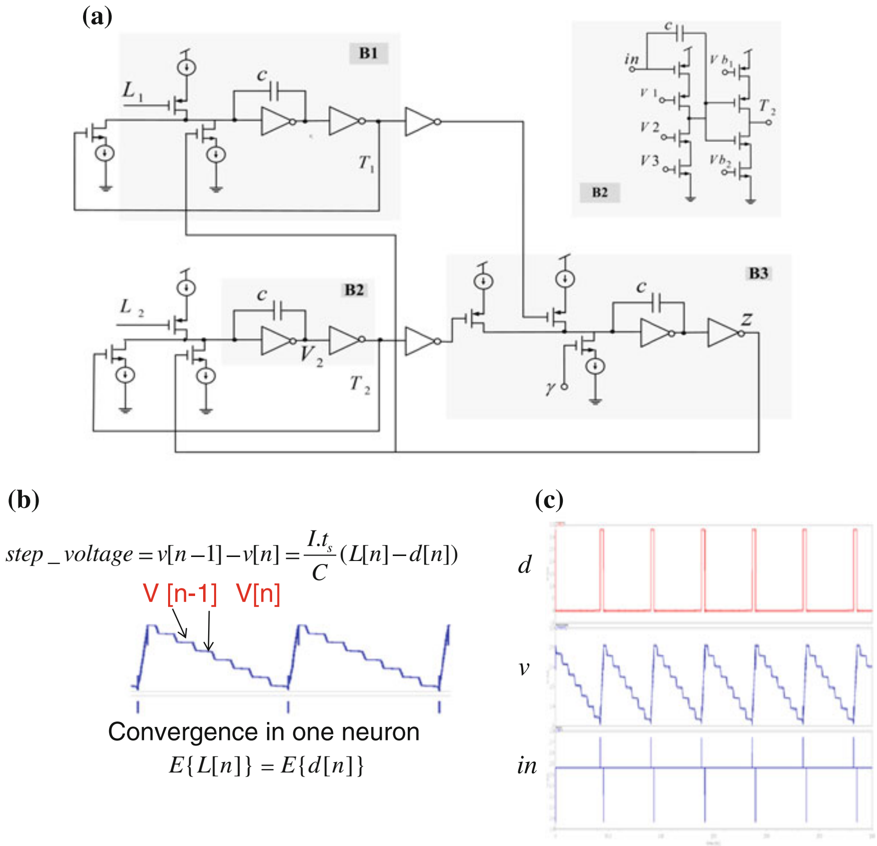


Fig. 2.6 **a** Schematic circuit of the IFMP model. **b** The membrane voltage of block *B1* and the convergence equation between the input and output rates for one neuron. **c** Output pulses of block *B1* (labled by *d* variable), the membrane voltage at the output of the integrator (labled by *v* variable), and input voltage of the integrator (labled by *in* variable)

2.4 IFMP: Chip Test Implementation

The dynamic characteristics of the IFMP module is verified in Matlab simulation (discussed in Sect. 2.2), cadence simulation and layout design on 0.5μ process. Figure 2.7b shows successfully the balance trend in the dynamic characteristics of one IFMP out of an array of 8 × 8 IFMP on a (1.5 × 1.5 mm) package. Experiments are implemented to test the dynamic characteristics for two inputs IFMP, when gamma changes between 0.2:0.6. In the experiments, the chip is biased to ensure the balance between the excitatory and inhibitory parts using National instrument (NI) data acquisition Input/ Output embedded systems. The applied biasing voltages are applied through Matlab program to initialize the ADC and

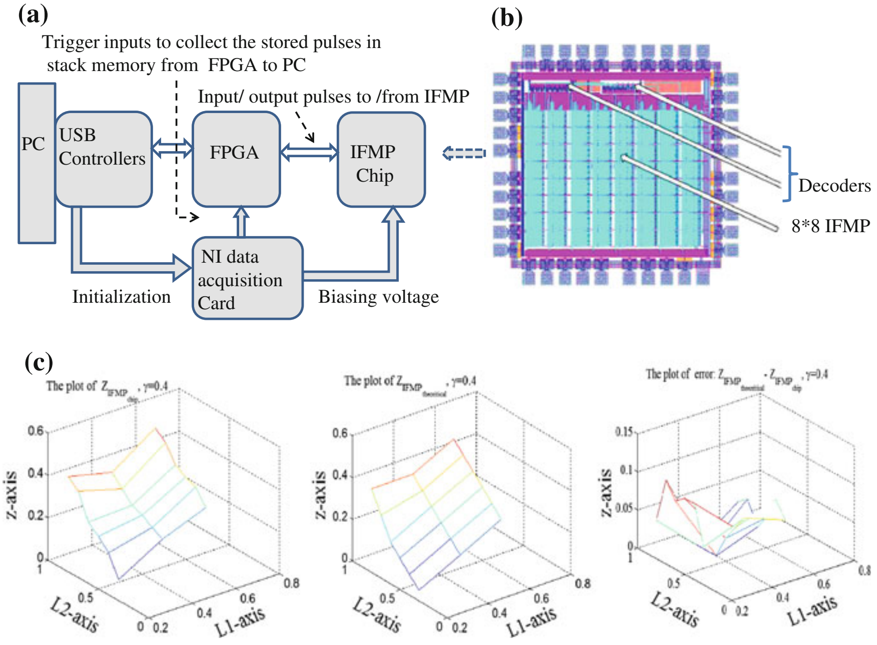

(a)
Trigger inputs to collect the stored pulses in stack memory from FPGA to PC
Input/ output pulses to /from IFMP
PC
USB Controllers
FPGA
IFMP Chip
NI data acquisition Card
Initialization
Biasing voltage
(b)
Decoders
8*8 IFMP
(c)
The plot of $Z_{IFMP_{chip}}$, γ=0.4
The plot of $Z_{IFMP_{theoretical}}$, γ=0.4
The plot of error: $Z_{IFMP_{theoretical}} - Z_{IFMP_{chip}}$, γ=0.4
Z-axis
L2-axis
L1-axis
0.6
0.4
0.2
0
1
0.5
0.8
0.15
0.1
0.05

Fig. 2.7 **a** System architecture to test the dynamic characteristics of IFMP chip **b**. **c** Dynamic characteristics that shows the output rate of IFMP unit and the error between the practical and theoritical rate

DAC converters of the NI card. The USB is supported with functions to communicate the input / outputs between the chip and FPGA. The inputs are applied as random pulses with specified rate. The pulses are generated using Verilog hardware description language (Verilog HDL). The HDL program is converted into bit file and configured into hardware components on FPGA using Xilinx Integrated Software Environment (ISE 9.2) software as shown in Fig. 2.7a. Samples of the results in Fig. 2.7b that shows the rate of two inputs IFMP output as theoretical and practical rate values when both inputs changed from 0.1:0.9 and the error between them when $\gamma = 0.4$.

2.5 IFMP Properties and Examples of Basic Computations

In Gu et al. [17], the properties were mentioned for analog margin propagation. In this work, the properties are tested in pulse computation mode. Experiments are implemented for both the theoretical and circuit pulse based margin propagation and shows the similarity between the two modes. The following margin propagation properties are verified in pulse mode.

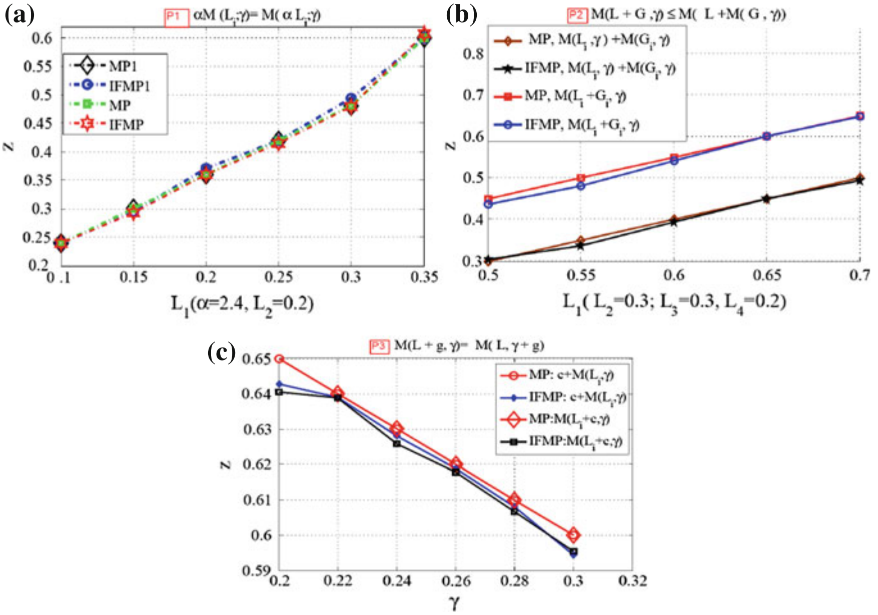


Fig. 2.8 Simulation of the right and left sides of theoritical and pulse margin properties of scaling ($P1$), superposition ($P2$) and offset ($P3$) in **a**, **b**, and **c** respectively. In the legend, the theoritical values of the two sides of the margin property are denoted as MP while the pulse rate values of margin denotes as the IFMP

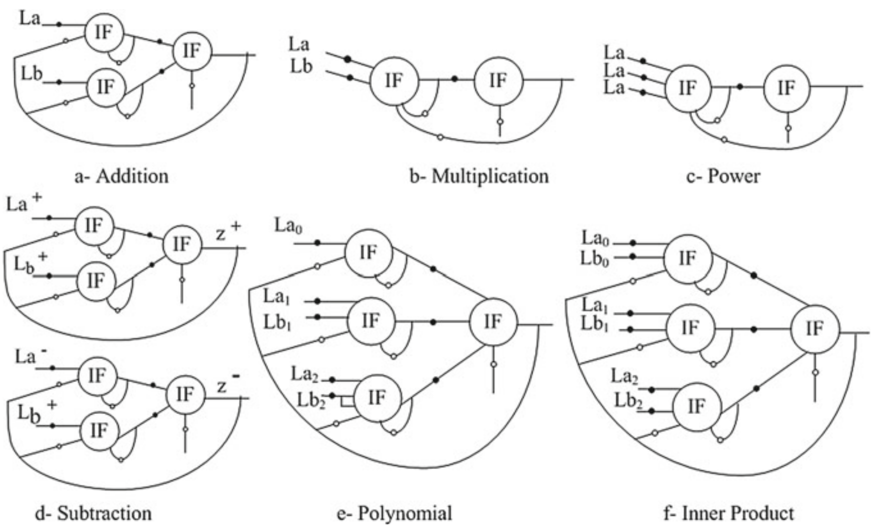
Property 1 (P_1, **Scaling Property**): For any $\alpha \in R, \alpha > 0$ and a set of series $\mathscr{L} = \{L_i\}, i = 1..N, M(\alpha\mathscr{L}, \alpha\gamma) = \alpha M(\mathscr{L}, \gamma)$.

Property 2 (P_2, **Superposition property):** Given two sets of scores $\mathscr{L}$ and $\mathscr{G}$ of size N with a well defined ordering, and if $\mathscr{L} + \mathscr{G}$ represent an element by element scalar addition, then $M(\mathscr{L} + \mathscr{G}, \gamma) \leq M(\mathscr{L}, \gamma) + M(\mathscr{G}, \gamma)$.

Property 3 (P_3, **Offset property):** Given a set of scores $\mathscr{L}$ of size N and a scalar $g \in \mathscr{R}$, then $M(\mathscr{L} + g, \gamma) = M(\mathscr{L}, \gamma) + g$. Property 3 states that if a constant offset to all the elements of input set leads to an equivalent offset in the output of the margin approximation function.

The computational properties above are applied successfully to the spiking networks based on the proposed IFMP model. It is shown in (a) to (c) of Fig. 2.8 the close match between the theoretical margin labled as MP in the figures and the pulse-mode margin labled (IFMP) for the above properties.

Before applying the model into algorithmic applications, It is necessary to show how the IFMP unit is used to approximate the basic computations such as addition,subtraction, multiplication, division, power, inner product and polynomial. The inputs to these units are probabilities and must be converted into logarithmic format then scaled the resultant negative values using the properties of scaling and



a- Addition

b- Multiplication

c- Power

d- Subtraction

e- Polynomial

f- Inner Product

Fig. 2.9 Basic computational architecture using IFMP

offset, discussed above, in order to keep the input values between zero and one. To retrieve the right computation values, the output of the computational function are scaled back in the reverse order of input scales as shown in Fig. 2.1. The basic computations are implemented using log-sum-exp, analog margin propagation and pulsed margin propagation as an approximated operation as well as the error for each of the above computations will be shown graphically for different values of input probability rates.

Addition: Let $f = a + b = e^{log(a+b)}$ then the exponent z can be represented in log-sum-exp and margin format as followings(see Fig. 2.9a)

$$z = log(e^{log(a)} + e^{log(b)}) = log(e^{L(a)} + e^{L(b)}) \tag{2.7}$$

$$z = M(L_a, L_b, \gamma) \tag{2.8}$$

where L stand for log function.

Multiplication: Let $f = a.b = e^{log(a.b)}$ then the exponent z can be represented in log-sum-exp and margin format as followings (see Fig. 2.9b)

$$z = log(e^{L(a)+L(b)}) \tag{2.9}$$

$$z = M(L_a + L_b, \gamma) \tag{2.10}$$

Division: Let $f = a/b$ and $c = 1/b$ then $f = ac$. We can represent the above as multiplication operation in IFMP representation as,

$$z = M(L_a + L_c, \gamma) \tag{2.11}$$

Power: Let $f = a_1a_2a_3 \ldots = e^{log(a_1a_2a_3\ldots)}$ then the exponent z can be represented in log-sum-exp and margin format as followings (see Fig. 2.9c)

$$z = log(e^{L(a_1)+L(a_2)\cdots}) \tag{2.12}$$

$$z = M(L_{a_1} + L_{a_2} + \cdots, \gamma) \tag{2.13}$$

Subtraction: In the subtraction function, we must use the differential form such that $f = a - b$ can be written as $f = (a^+ - a^-) - (b^+ - b^-) = (a^+ + b^-) - (b^+ + a^-)$

$$f = e^{log(a^+ + b^-)} - e^{log(b^+ + a^-)} \tag{2.14}$$

$$f = e^{L(e^{La^+} + e^{Lb^-})} - e^{L(e^{Lb^+} + e^{La^-})} \tag{2.15}$$

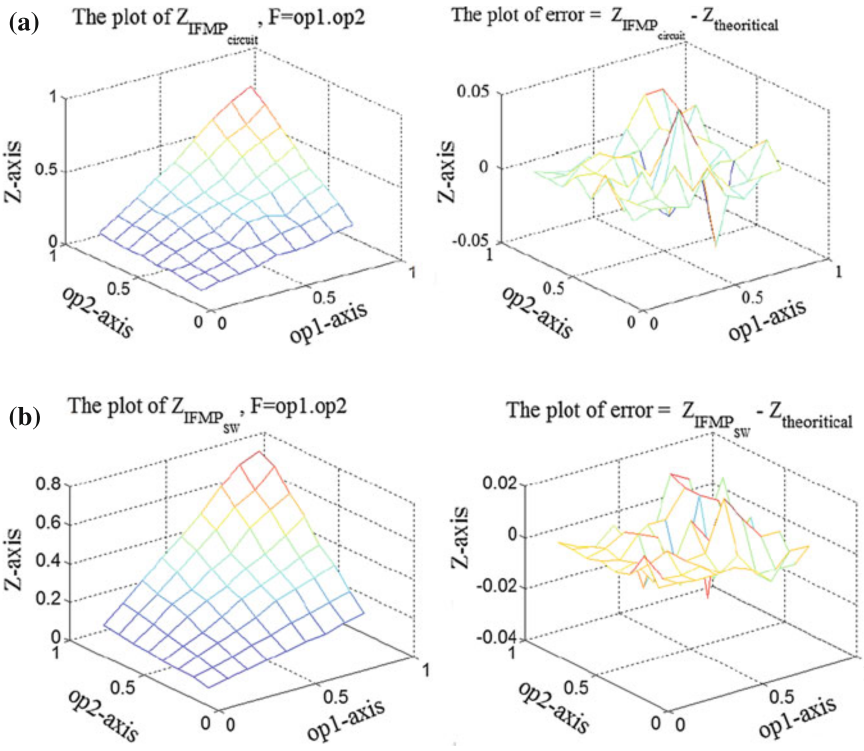


Fig. 2.10 Pulse computational for Multiplication function and the error compared with the theoritical IFMP using circuit simulation a and sotware simulation **b**

The following two exponents z^{+} and z^{-} are then evaluated using two IFMP's as shown in Fig. 2.9d such that

$$z^{+} = M(L_{a^{+}}, L_{b^{-}}, \gamma). \tag{2.16}$$

and

$$z^{-} = M(L_{b^{+}}, L_{a^{-}}, \gamma). \tag{2.17}$$

Polynomial: Let $f = a_0 + a_1 b_1 + a_2 b_2^2 + a_3 b_3^3$ then the exponent z can be represented in log-sum-exp and margin format as followings (see Fig. 2.9e)

$$z = log(e^{L_{a_0}} + e^{L_{a_1} + L_{b_1}} + e^{L_{a_2} + 2L_{b_2} + \cdots}) \tag{2.18}$$

$$z = M(L_{a_0}, L_{a_1} + L_{b_1}, L_{a_2} + 2L_{b_2} \cdots, \gamma) \tag{2.19}$$

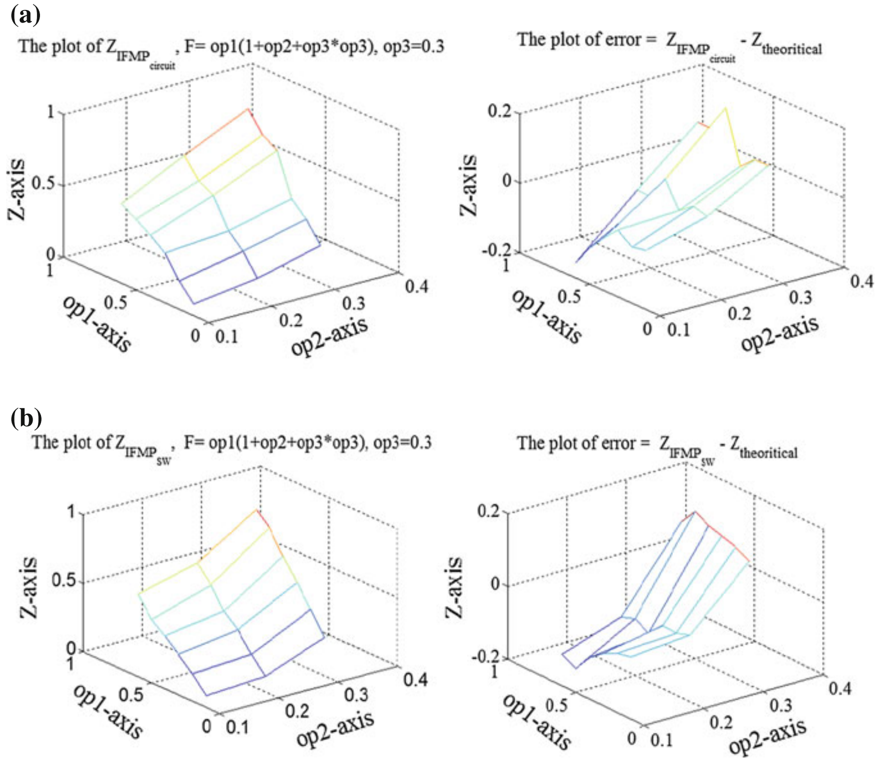


Fig. 2.11 Pulse computational for inner product function and the error compared with the theoritical IFMP using circuit simulation **a** and sotware simulation **b**

Inner Product: Let $f = a_0b_0 + a_1b_1 + a_2b_2 + ...$ then the exponent z can be represented in log-sum-exp and margin format as followings (see Fig. 2.9f)

$$z = log(e^{L_{a_0}+L_{b_0}} + e^{L_{a_1}+L_{b_1}} + e^{L_{a_2}+L_{b_2}} + \cdots) \tag{2.20}$$

$$z = M(L_{a_0} + L_{b_0}, L_{a_1} + L_{b_1}, L_{a_2} + L_{b_2} \cdots, \gamma) \tag{2.21}$$

We are showing samples of results for the pulse computations in circuit simulation and compare them with theoretical values for two functions. In the left part of Figs. 2.10 and 2.11, the rates of output pulses are evaluated using circuit simulation and matlab simulations of IFMP modules for multiplication and inner product respectively. The input operands are represented by the rates of random pulses for pre-assigned time period ($1,500\, samples \times 0.2$ ms). The right part of the figure represents the error between the theoretical LSE rates and pulse based IFPM rates in the two cases "circuit simulation and matlab simulations".

2.6 Conclusion

We designed a novel pulse computational module to implement basic functions used in many algorithms. For this purpose, a scalable IFMP computational module is analysed, mapped and verified into an analog design circuit on a standard $0.5\mu m$ process. The IFMP properties are verified in simulation and the computation functions are verified and tested in simulation and analog circuit hardware. The importance of this module is to map a pulsed mode algorithm into low power silicon circuits as an attempt to implement signal processing in the neuromorphic applications. The IFMP network can be expanded to implement generalized machine learning architectures like Hidden Markov Models "HMM". The above applications are already verified using IFMP module in both Matlab and cadence simulation. In Hindo [18], the above has been verified by implementing HMM sequence recognition task in analog circuit by using pulse based computation modules IFMP. These applications are to be tested and verified in hard ware using the designed chip as a current work. Achieving MP based computation in spike-domain in scalable way opens the possibility of implementing hybrid neuromorphic architectures where large-scale machine learning algorithms can now be integrated on spiking neural-networks. Even though this chapter focused on MP computing using rate-based encoding, we believe that the IFMP principle can also be extended to other spike-encoding techniques like time-to-spike encoding or variable spike-encoding techniques, both of which will be the subject of future research in this area.

References

1. Mead CA (1989) Analog VLSI and neural systems. Addison-Wesley, Boston
2. Mahowald M (1992) VLSI analogs of neuronal visual processing: a synthesis of form and function. Technical report. California Institute of Technology, Pasadena
3. Boahen K (2005) Neuromorphic microchips. Sci Am 292(5):56–63
4. Liu S-C, Delbruck T (2010) Neuromorphic sensory systems. Curr Opin Neurobiol 20:1–8
5. Hindo T, Chakrabartty S (2012) Noise-exploitation and adaptation in neuromorphic sensors. In: Proceeding of SPIE, bioinspiration, biomimetics, and bioreplication, vol 8339, March 2012
6. Chakrabartty S, Cauwenberghs G (2004) Margin normalization and propagation in analog vlsi. In: ISCAS (1)'04, pp 901–904
7. Loeliger HA (2004) An introduction to factor graphs. Signal Process Mag IEEE 21(1): 28–41. http://dx.doi.org/10.1109/MSP.2004.1267047
8. Kong C, Chakrabartty S (2007) Analog iterative ldpc decoder based on margin propagation. Circuits Syst II Express Briefs IEEE Trans 54(12):1140–1144
9. Gu M, Chakrabartty S (2012) Synthesis of bias-scalable cmos analog computational circuits using margin propagation. Circuits Syst I Regul Pap IEEE Trans 59(2):243–254
10. Izhikevich EM (2003) Simple model of spiking neurons. IEEE Trans Neural Networks 14:1569–1572
11. Gerstner W, Kistler WM (2002) Spiking neuron models: single neurons, populations, plasticity, 1st edn. Cambridge University Press, Cambridge
12. Segee B (1999) Methods in neuronal modeling: from ions to networks. Comput Sci Eng 1:81
13. Indiveri G et al (2011) Neuromorphic silicon neuron circuits. Frontiers Neurosci 5:1–23
14. Sarpeshkar R (1998) Analog versus digital: extrapolating from electronics to neurobiology. Neural Comput 10(7):1601–1638
15. Li Y, Shepard K, Tsividis Y (2005) Continuous-time digital signal processors. Asynchronous circuits and systems, ASYNC 2005. In: Proceedings of 11th IEEE international symposium, pp 138–143, March 2005
16. Hindo T (2012) Scalable pulsed computational module using integrate and fire structure and margin propagation algorithm. In: Proceedings of the world congress on engineering and computer science, vol 2. WCECS 2012, pp 860–865, Oct 2012
17. Gu M, Misra K, Radha H, Chakrabartty S (2009) Sparse decoding of low density parity check codes using margin propagation. In: Global telecommunications conference, vol 2009. GLOBECOM 2009. IEEE, pp 1–6, 30 Dec 2009
18. Hindo T (2012) An asynchronous, time-domain analog hidden markov circuit based on integrate and fire margin propagation. In: ASME: proceeding of 5th conference on computer and electrical engineering, Oct 2012

Chapter 3
Hardware Implementation of Microprogrammed Controller Based Digital FIR Filter

Syed Manzoor Qasim and Mohammed S. BenSaleh

Abstract Digital finite-impulse response (FIR) filter is the fundamental processing element of many digital signal processing (DSP) systems, ranging from wireless communications to image and video processing. The microarchitecture of digital FIR filter consists of a datapath and a control unit. The datapath is the computational engine of FIR filter and mainly consists of adders, multipliers and delay elements. Several techniques have been proposed in the existing literature to implement digital FIR filters in hardware using field programmable gate array (FPGA). In this chapter, hardware implementation of a parallel digital FIR filter architecture using a novel microprogrammed controller is presented. The main advantage of the microprogrammed controller is its flexibility in modifying the microprogram stored in ROM based control memory. To demonstrate the proposed technique, a 4-tap parallel FIR filter is implemented using Virtex-5 FPGA. The proposed FIR filter is coded in VHDL using top-down hierarchical design methodology. Performance evaluation is done based on the implementation results obtained through FPGA synthesis tools. The designed 4-tap FIR filter utilizes minimal area leaving bulk of the FPGA resources to implement other parallel processors on the same device. The design can be easily modified to implement higher-order and high speed FIR filters which are commonly used in video and image processing applications.

Keywords Digital filter · Finite-impulse response (FIR) filter · FPGA · Hardware · Implementation · Microprogrammed controller

S. M. Qasim (✉) · M. S. BenSaleh
King Abdulaziz City for Science and Technology (KACST), National Center
for Electronics, Communications and Photonics (ECP), Riyadh 11442,
Kingdom of Saudi Arabia
e-mail: mqasim@kacst.edu.sa

M. S. BenSaleh
e-mail: mbensaleh@kacst.edu.sa

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_3,

3.1 Introduction

Digital filters are one of the most widely used building blocks of many digital signal processing (DSP) systems. They are most commonly used in signal, image and video processing applications. Digital filters are an important class of linear time-invariant (LTI) systems designed for filtering out undesirable parts (random noise) from the signal, spectral shaping, motion estimation, noise reduction and channel equalization among many other applications. Finite impulse response (FIR) and infinite impulse response (IIR) are two such digital filters used in different applications. The choice between an FIR filter and an IIR filter depends on the application requirements [1]. However, FIR filters are more commonly used because of their absolute stability and linear phase properties.

Adders, multipliers and delay elements are the main components used in the implementation of digital FIR filters. These components are arranged and interconnected in different ways based on the architecture of the FIR filter [2]. Basically, FIR filter performs a linear convolution on a window of N data samples which can be mathematically expressed as follows:

$$y(k) = \sum_{n=0}^{N-1} w(n) \cdot x(k-n) \tag{3.1}$$

For example, the difference equation for 4-tap FIR filter can be written as

$$y(k) = w_0 \cdot x(k) + w_1 \cdot x(k-1) + w_2 \cdot x(k-2) + w_3 \cdot x(k-3) \tag{3.2}$$

A direct form implementation of an FIR filter can be readily developed from the convolution sum as shown in Fig. 3.1. It is called direct form because the multiplier coefficients are obtained directly from the filter transfer function. Direct form FIR filters are also known as tapped delay line or transversal filters. The size of the FIR filter is determined by the number of coefficients. As can be seen in Fig. 3.1, N-tap FIR filter consist of N delay elements, N multipliers and N-1 adders or accumulators. The impulse response of the FIR filter can be directly inferred from the tap coefficients w_n.

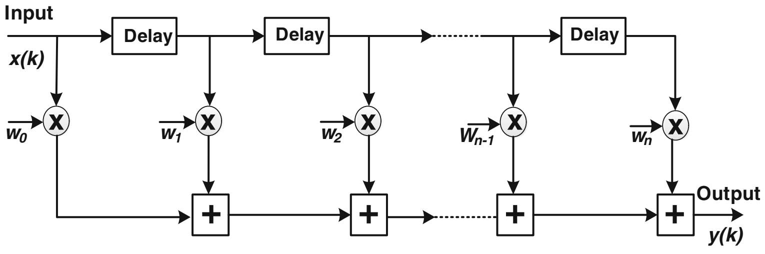


Fig. 3.1 Block diagram of direct form N-tap FIR filter

Several techniques for the hardware implementation of digital FIR filter using Field Programmable Gate Arrays (FPGAs) have been reported in the literature [3–5]. However, design and FPGA implementation of digital FIR filter using microprogrammed controller [6–8] has not been reported in the open literature. The main advantage of using a microprogrammed controller is that the state machine can adapt to changing algorithms by changing a bit pattern in the control memory which has no impact on FPGA logic resources or timing. An added benefit of using microprogrammed approach is often a more structured organization of the controller [6].

FPGAs have improved considerably in logic density, functionality and speed, thus making them ideal for System-on-Chip (SoC) designs for wide range of applications. Today, FPGAs are large and fast enough for use in multimillion gate DSP system designs. This chapter presents the details of the proposed technique using an example of 4-tap parallel digital FIR filter. 4-tap FIR filters are normally used as vertical scalers in video processing units. The design can be easily modified to implement higher order FIR filters. This chapter is an extended version of the conference paper presented in [9]. The chapter has been extended to provide more background information, implementation details and new results.

The rest of the chapter is organized as follows: Sect. 3.2 presents the FIR filter architecture with details of datapath and microprogrammed controller discussed in Sects. 3.3 and 3.4 respectively. Hardware implementation using FPGA and simulation results are further presented in Sects. 3.5 and 3.6 respectively. Section 3.7 concludes the chapter with some directions for future work.

3.2 FIR Filter Architecture

The proposed FIR filter architecture can be partitioned into two main blocks which are datapath and control unit.

The top level design of a generic FIR filter with the integrated datapath and control unit is shown in Fig. 3.2. The datapath is the computational engine of FIR filter and the control unit orchestrates the operation of datapath.

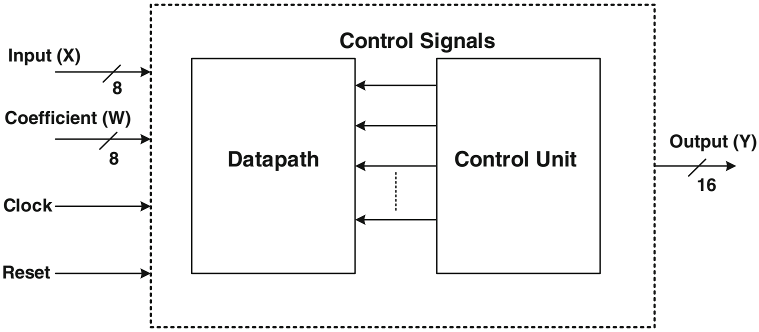


Fig. 3.2 Top level design of generic FIR filter

3.3 Datapath Microarchitecture

The datapath microarchitecture may vary depending on the application and requirement. Parallel and sequential datapath architectures are the most commonly used ones.

The datapath microarchitecture of 4-tap parallel FIR filter as shown in Fig. 3.3 consists of the following sub modules [9, 10]:

- Four 8-bit data registers
- One 2-to-4 decoder

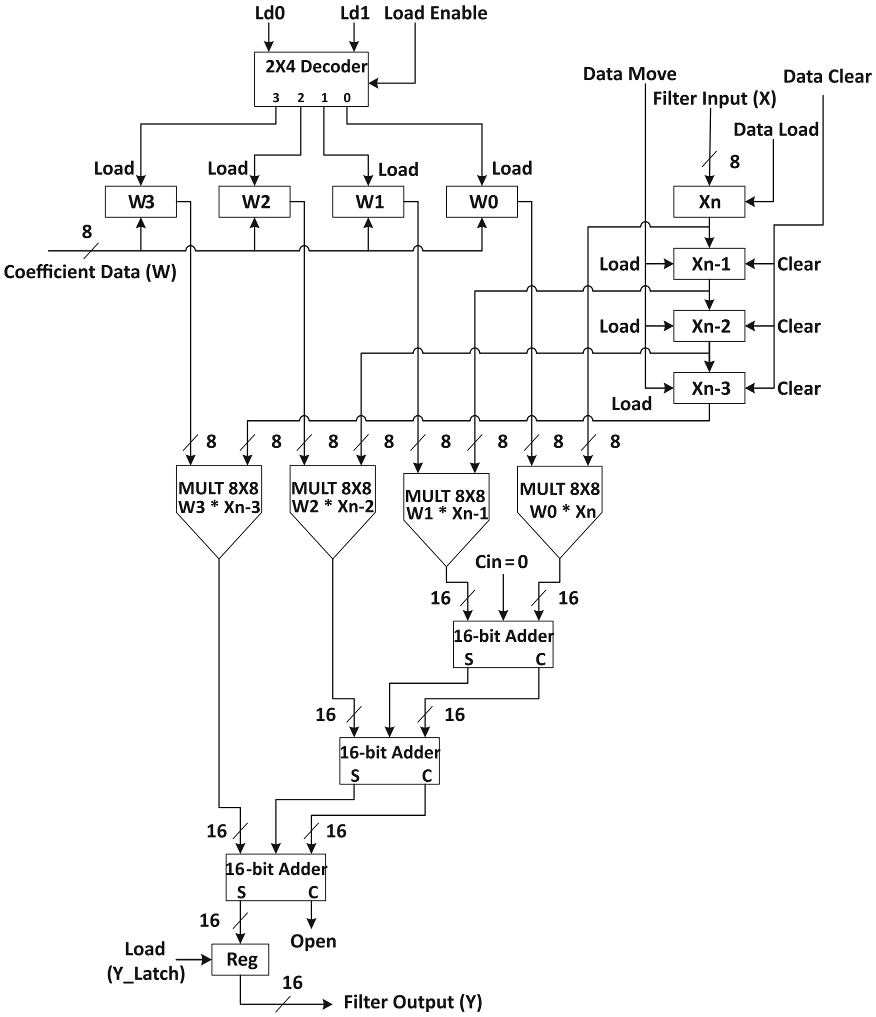


Fig. 3.3 Datapath microarchitecture of 4-tap FIR filter

- Four 8-bit coefficient registers
- Four multipliers (8 × 8)
- Three 16-bit adders
- One 16-bit register for latching the output

Each sub modules are coded in VHDL and finally integrated to obtain the complete datapath. The control signals generated by the microprogrammed controller for this datapath are fed to different sub modules for proper operation of the FIR filter.

3.4 Microprogrammed Controller

There are several methods to design the control unit of FIR filter, such as hard-wired controller and microprogrammed controller. In this chapter, the control logic of FIR filter is implemented using microprogrammed controller [8, 10]. The main advantage of the microprogrammed controller is its flexibility to modify the microprogram in the ROM based control memory [11, 12]. This makes the design of higher order FIR filter much easier.

As shown in Fig. 3.4, microprogrammed controller consists of two main parts. The first part is responsible for addressing microinstructions kept in the control memory and the second part is used to hold and generate microinstructions for the datapath. The sequence of operations listed in Table 3.1 is followed to generate the FIR filter output.

Table 3.1 presents the control information for the parallel architecture. The word stored in the control memory consist of three parts: 1-bit control SEL (CS) signal for signaling the microprogram counter (MPC) either to count or to load external branch address, the next four bit represents the branch address and the remaining bits represent the control function for the datapath.

For the given architecture, the microprogrammed controller generates seven control signals (12-bit microcode) for the FIR filter datapath. These control signals are then fed to different sub modules of the FIR filter datapath for proper operation. At the heart of the microprogrammed controller is a ROM based control memory. The MPC holds the address of the next microinstruction to be executed and it is incremented after each microinstruction fetch [10].

As can be seen in Table 3.1, the FIR filter tap coefficient registers are loaded with data depending on load enable (LoadEn) signal and the decoder signals (Ld1 and Ld0). After loading the tap coefficient registers, all the input registers are cleared by asserting data clear (Dclear) signal high and then the filter input data is entered into first data register after the data load (Dload) signal is asserted high. The output (filtered data) is available only after the latch output (Ylatch) signal is asserted high. The process is continued for the remaining registers only after the data move (Dmove) signal is asserted high.

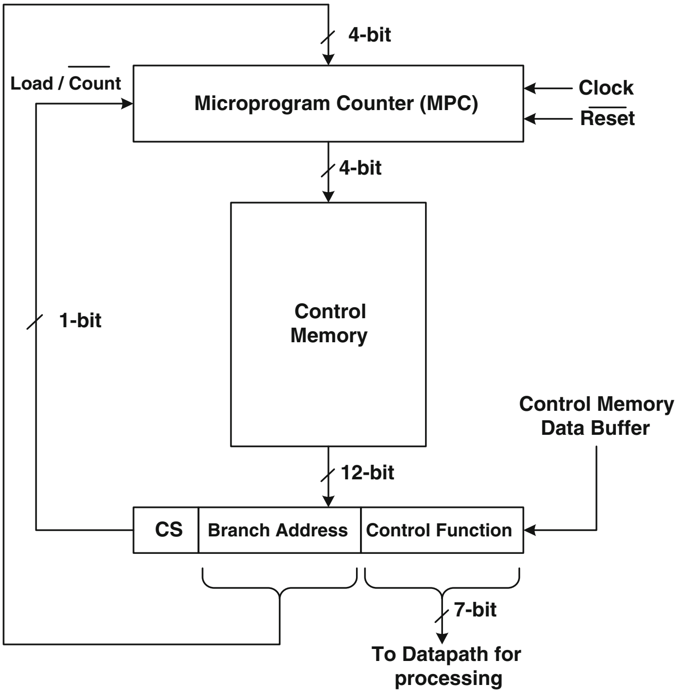


Fig. 3.4 Microprogrammed Controller

Table 3.1 Microprogram control signals for the datapath

No.	CS	Branch address	Control functions							Comments
			LoadEn	Ld1	Ld0	Dclear	Dload	Dmove	Ylatch	
1	0	0 0 0 0	1	0	0	0	0	0	0	Load w_0
2	0	0 0 0 0	1	0	1	0	0	0	0	Load w_1
3	0	0 0 0 0	1	1	0	0	0	0	0	Load w_2
4	0	0 0 0 0	1	1	1	1	0	0	0	Load w_3
5	0	0 0 0 0	0	0	0	0	1	0	0	Load data
6	0	0 0 0 0	0	0	0	0	0	1	0	Move data
7	0	0 0 0 0	0	0	0	0	0	0	1	Latch output
8	1	0 1 0 0	0	0	0	0	0	0	0	Go to step 5

3.5 Hardware Implementation Using FPGA

The parallel FIR filter is designed and simulated using VHDL. The proposed architecture is synthesized and implemented in Virtex-5 (5vlx50tff1136-3) FPGA using Xilinx XST and Synplify Pro synthesis tools. A comparison of the FPGA resource utilization results generated by Xilinx ISE and Synplify pro tools are

Table 3.2 FPGA resource utilization using Xilinx XST synthesis tool

FPGA resources	Number used/available	Utilization (%)
Slice registers	72 out of 28800	1
Slice LUTs	96 out of 28800	1
Occupied slices	43 out of 7200	1
Number of IOBs	34 out of 480	7
DSP48Es	4 out of 48	8
Minimum period (ns)	9.363	–
Maximum frequency (MHz)	106.803	–
Total runtime (s)	87	–

Table 3.3 FPGA resource utilization using Synplify Pro synthesis tool

FPGA resources	Number used/available	Utilization (%)
Slice registers	52 out of 28800	1
Slice LUTs	101 out of 28800	1
Occupied slices	44 out of 7200	1
Number of IOBs	34 out of 480	7
DSP48Es	3 out of 48	8
Minimum period (ns)	11.544	–
Maximum frequency (MHz)	86.6	–
Total runtime (s)	26	–

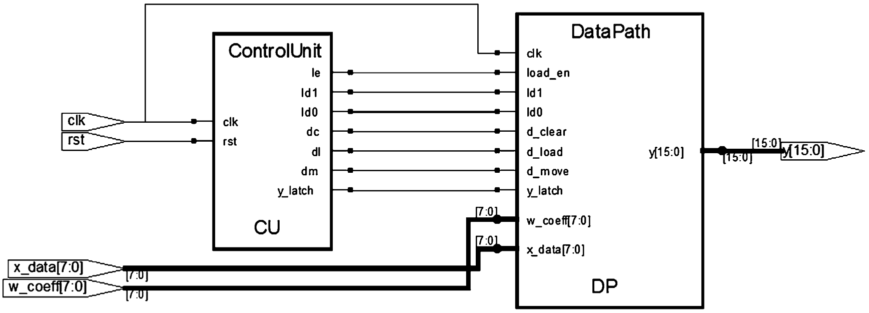
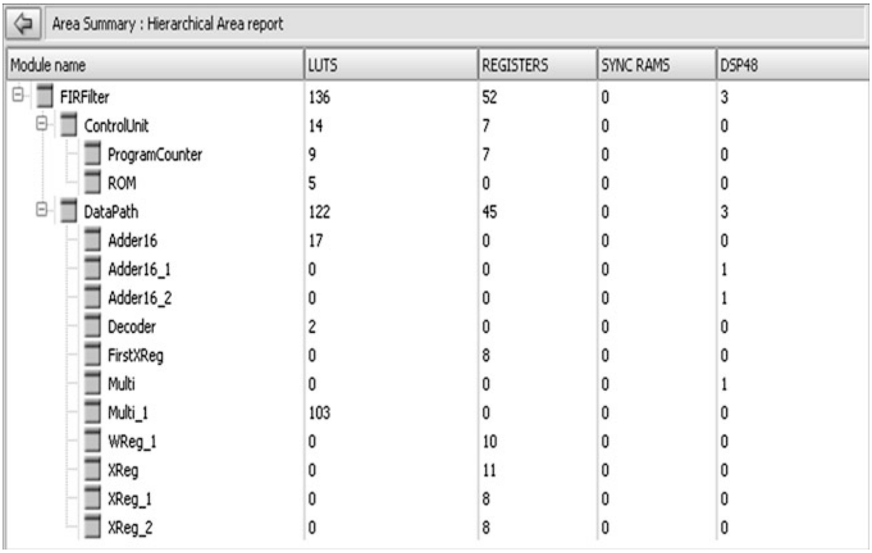


Fig. 3.5 Top level RTL schematic

summarized in Tables 3.2 and 3.3 respectively. The top level RTL schematic of the FIR filter generated by Synplify Pro tool is shown in Fig. 3.5.

Figure 3.6 presents the area summary in the form of hierarchical area report. The designed FIR filter architecture has basically two levels of hierarchy: control unit and datapath. The control unit is further decomposed into lower level of hierarchy comprising of program counter and control memory (ROM). The datapath is also decomposed into another lower level of hierarchy consisting of several combinational and sequential logic elements which are implemented using LUTs, slice registers and DSP48E blocks of Virtex-5 FPGA.



Area Summary : Hierarchical Area report

Module name	LUTS	REGISTERS	SYNC RAMS	DSP48
FIRFilter	136	52	0	3
ControlUnit	14	7	0	0
ProgramCounter	9	7	0	0
ROM	5	0	0	0
DataPath	122	45	0	3
Adder16	17	0	0	0
Adder16_1	0	0	0	1
Adder16_2	0	0	0	1
Decoder	2	0	0	0
FirstXReg	0	8	0	0
Multi	0	0	0	1
Multi_1	103	0	0	0
WReg_1	0	10	0	0
XReg	0	11	0	0
XReg_1	0	8	0	0
XReg_2	0	8	0	0

Fig. 3.6 Hierarchical area report (using Synplify Pro)

The designed FIR filter operates at a maximum clock frequency of 106.803 MHz and 86.6 MHz using Xilinx XST and Synplify Pro tool respectively. 4-tap FIR filter consumes a small area out of the entire FPGA real estate leaving plenty of resources for implementing other parallel processors [13]. The results demonstrate that faster synthesis runtime and area efficient results are achieved with Synplify Pro as compared to Xilinx XST technology.

3.6 Simulation Results

Three different test cases are used for testing the designed FIR filter circuit. The tap coefficients are chosen randomly with an objective to provide something that is observable at the FIR filter output. These taps could be changed depending on the requirement of the application [14]. The functionality of the parallel FIR filter is verified through simulation using Xilinx ISE simulator.

Figure 3.7 presents a snapshot of the simulation results for the microprogrammed controller. Figures 3.8, 3.9 and 3.10 presents the simulation waveform of the datapath unit for three different test cases as listed in Table 3.4. Finally, the datapath unit and microprogrammed controller are integrated together to demonstrate the simulation results of 4-tap FIR filter for each test case. The simulation waveforms of the top level FIR filter for each test case are presented in Figs. 3.11, 3.12, and 3.13 respectively.

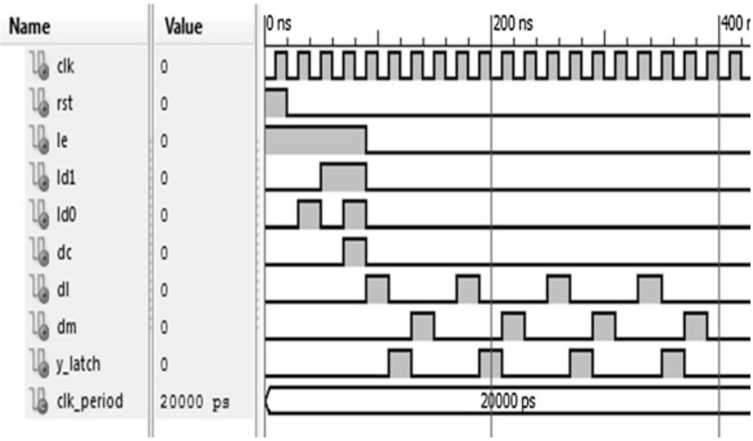


Fig. 3.7 Microprogrammed controller simulation waveform

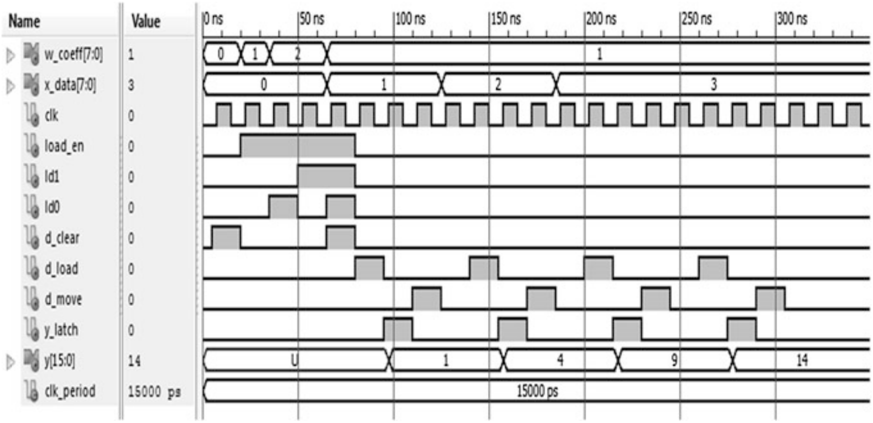


Fig. 3.8 Datapath simulation waveform (test case 1)

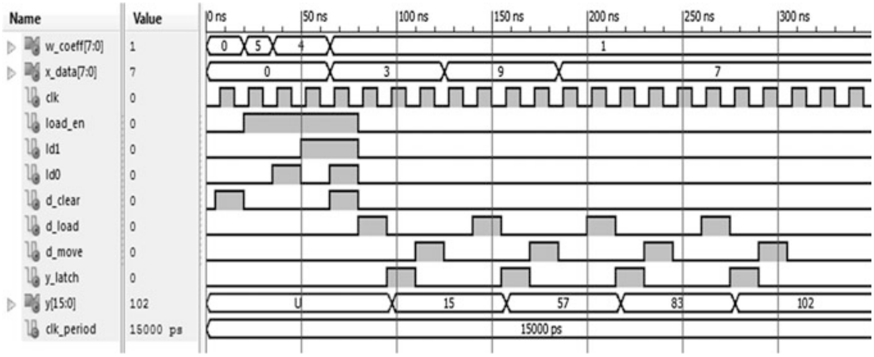


Fig. 3.9 Datapath simulation waveform (test case 2)

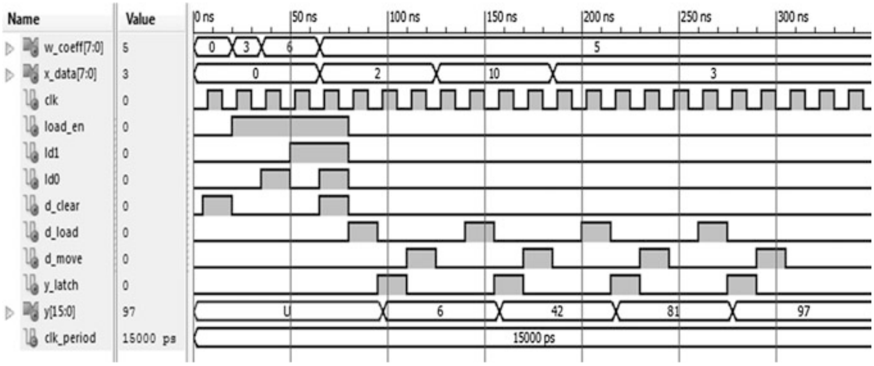


Fig. 3.10 Datapath simulation waveform (test case 3)

Table 3.4 Simulation test cases

Test cases	Input data (X)	Tap coefficients (W)	Output data (Y)
1	[1, 2, 3, 3]	[1, 2, 2, 1]	[1, 4, 9, 14]
2	[3, 9, 7, 7]	[5, 4, 4, 1]	[15, 57, 83, 102]
3	[2, 10, 3, 3]	[3, 6, 6, 5]	[6, 42, 81, 97]

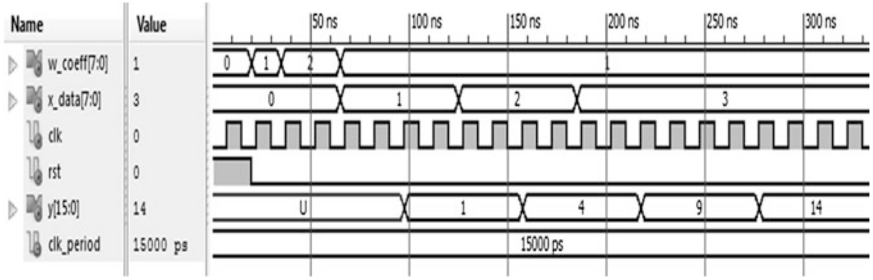


Fig. 3.11 FIR filter simulation waveform (test case 1)

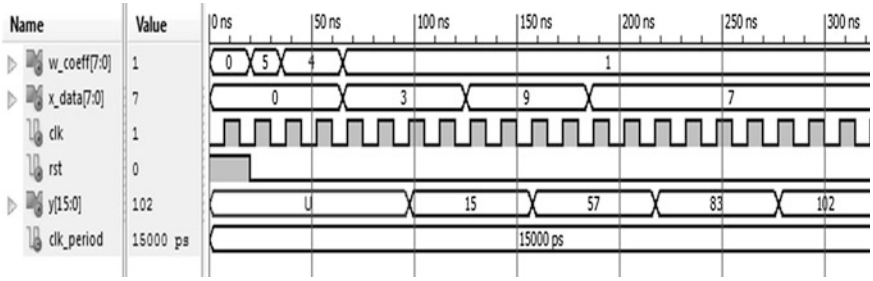


Fig. 3.12 FIR filter simulation waveform (test case 2)

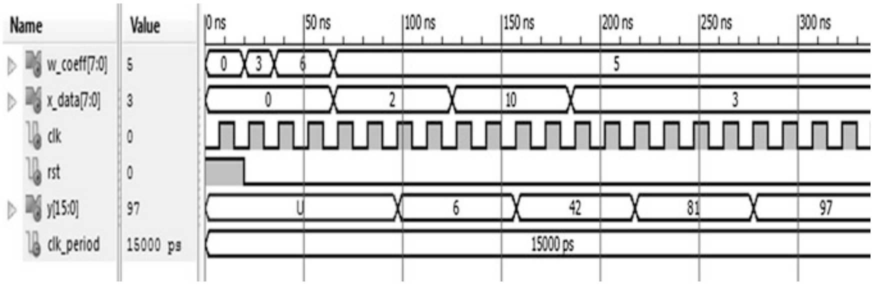


Fig. 3.13 FIR filter simulation waveform (test case 3)

3.7 Conclusion and Future Work

Hardware implementation of microprogrammed controller based 4-tap digital FIR filter using FPGA is presented in this chapter. The microprogrammed controller is used for controlling the operation of the filter. A parallel datapath architecture utilizing four multipliers and three adders along with other building blocks is used to demonstrate the proposed technique. Performance evaluation is done by synthesizing and implementing the design in Virtex-5 FPGA using Xilinx XST and Synplify pro tools. FPGA based hardware implementation results demonstrate that the design can operate at a maximum clock frequency of 106.803 and 86.6 MHz using Xilinx XST and Synplify Pro tools respectively. The results also demonstrate that faster synthesis runtime and area efficient results are obtained with Synplify Pro as compared to Xilinx XST technology. In both the cases, the design consumes a small area out of the entire FPGA real estate leaving plenty of resources for implementing other parallel processors on the same device.

Future work would focus on developing a stand-alone generic FIR filter intellectual property (IP) core based on the proposed microprogrammed controller technique. It is also envisioned to develop an equivalent sequential architecture of the FIR filter. Different optimization techniques will be applied to the design and a comparison of parallel and sequential architecture for speed, area and power will be done.

Acknowledgments This work was supported in part by King Abdulaziz City for Science and Technology (KACST) under Project Ejaz (IRU No. 31/513). The authors would like to thank the reviewers for their valuable comments and suggestions to improve the quality of this chapter.

References

1. Parhi KK (1999) VLSI digital signal processing systems: design and implementation. Wiley, New York
2. Khan SA (2011) Digital design of signal processing systems: a practical approach. Wiley, West Sussex

3. Nekoei F, Kavian YS, Strobel O (2010) Some schemes of realization digital FIR filters on FPGA for communication applications. In: Proceedings of 20th international Crimean conference on microwave and telecommunication technology (CriMiCo), pp 616–619
4. Zhou Y, Shi P (2011) Distributed arithmetic for FIR filter implementation on FPGA. In: Proceedings of IEEE international conference on multimedia technology (ICMT), pp 294–297
5. Meyer-Baese U, Botella G, Romero DET, Kumm M (2012) Optimization of high speed pipelining in FPGA-based FIR filter design using genetic algorithm. In: Proceedings of SPIE 8401
6. Bomar BW (2002) Implementation of microprogrammed control in FPGAs. IEEE Trans Industr Electron 49(2):415–422
7. Barkalov A, Titarenko L (2008) Logic synthesis for compositional micro-program control units. Springer, Berlin
8. Wiśniewski R, Barkalov A, Titarenko L, Halang W (2011) Design of microprogrammed controllers to be implemented in FPGAs. Int J Appl Math Comput Sci 21(2):401–412
9. BenSaleh MS, Qasim SM, Bahaidarah M, AlObaisi H, AlSharif T, AlZahrani M, AlOnazi H (2012) Field programmable gate array realization of microprogrammed controller based parallel digital FIR filter architecture. In: Lecture notes in engineering and computer Science: proceedings of the world congress on engineering and computer science, WCECS 2012, 24–26 October, 2012 San Francisco, USA, pp 828–831
10. Rafiquzzaman M (2005) Fundamentals of digital logic and microcomputer design. Wiley, New Jersey
11. Barkalov AA, Titarenko LA, Efimenko KN (2011) Optimization of circuits of compositional microprogram control units implemented on FPGA. Cybern Syst Anal 47(1):166–174
12. Wisniewski R, Wisniewska M, Wegrzyn M, Marranghello N (2011) Design of microprogrammed controllers with address converter implemented on programmable systems with embedded memories. In: Proceedings of 9th IEEE east-west design and test symposium (EWDTS), pp 123–126
13. Amos D, Lesea A, Richter R (2011) FPGA-based prototyping methodology manual: best practices in design-for-prototyping. Synopsys Press, California
14. Mukherjee N, Rajski J, Tyszer J (2001) Testing schemes for FIR filter structures. IEEE Trans Comput 50(7):674–688

Chapter 4
Drude-Lorentz Model of Semiconductor Optical Plasmons

Mohamed Eldlio, Franklin Che and Michael Cada

Abstract Theoretical solutions are obtained for the propagation of electromagnetic waves at optical frequencies along a semiconductor/dielectric interface when losses are taken into account in the form of a complex dielectric function. A combination method for the dielectric function, comprised of the best features of the Drude and Lorentz models, is herein proposed. By including the loss term in both models, we were able to obtain numerical solutions for the Plasma dispersion curve of the semiconductor/dielectric interface. The surface plasmon waves, when excited, become short wavelength waves in the Optical frequency or THz region. A silicon/air structure was used as our semiconductor/dielectric material combination, and comparisons were made to optical plasmons generated without losses. Our initial numerical calculation results show enormous potential for use in several applications.

Keywords Drude-Lorentz model · Optical frequency · Optical plasmon · Plasma dispersion · Semiconductor optical plasmons · Surface plasmon polaritons

4.1 Introduction

The fundamental optical excitation that is confined to a metal/dielectric interface is the Surface Plasmon Polariton (SPP), as described by Ritchie [1]. The term SPP comes from coupled modes, which can be used to confine light and increase the

M. Eldlio (✉) · F. Che · M. Cada
Department of Electrical and Computer Engineering, Dalhousie University,
Halifax, Nova Scotia B3H4R2, Canada
e-mail: meldlio@dal.ca

F. Che
e-mail: franklin.che@dal.ca

M. Cada
e-mail: michael.cada@dal.ca

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_4,

electromagnetic fields at an interface between two media, of which at least one is conducting [2–5]. Plasmonics in a semiconductor is taking an increasingly prominent role in the design of future silicon-based optoelectronic chips [6].

Optical plasmons have been shown to have many applications [2, 3] and are generally excited using metal/dielectric interfaces due to the high concentration of charge carriers in metals. SPPs in semiconductor/dielectric interfaces have recently received considerable interest, and the use of a semiconductor/dielectric interface to support optical plasmons has been numerically shown in [7], albeit without the inclusion of losses. We therefore wish to take this a step further by including the loss contribution in the SPP dispersion relation. The loss is introduced through the complex dielectric function of the semiconductor. Little attention has thus far been paid to this phenomenon because it is generally very difficult to deal with. The SPP's dispersion and the resonance frequency depend on the interface configuration [2, 7].

Unlike metals, the semiconductor permittivity theory can be extremely complex, since it depends on the doping concentration of the semiconductor, which also determines the number of bound and free charge carriers in the material. In a semiconductor, plasma frequency can be determined by the effective carrier mass as well as the doping concentration. Two of the more difficult tasks are to confine light and increase the electromagnetic fields near the interfaces. The losses are related to the Plasma dispersion; however, they affect the shape of the plasmon dispersion curve near the plasma frequency.

Our goal is to develop a theoretical treatment to find a suitable model for the dielectric function of semiconductors. Starting with the Drude model, which is commonly used to describe the dielectric function of metals, we seek to modify and adapt it for semiconductors by adding the Lorentz model.

We proceed by solving Maxwell's equation for the interface between the dielectric and the semiconductor, and using the dielectric function described by the Drude-Lorentz model to obtain the dispersion relation.

This paper is organized as follows: A brief theory and background of models describing dielectric permittivity is addressed in Sect. 4.2. In Sect. 4.3, numerical results are presented and discussed, and Sect. 4.4 summarizes the results and draws conclusions. This is an extended and revised work of an earlier published conference paper [8].

4.2 Theoretical Analysis

In this section, we briefly review different models, one of which is selected for our approach. The Drude model, Lorentz model, and a combination of both models (i.e., Drude-Lorentz model) are presented and discussed. Starting with a semiconductor/dielectric interface, we seek to obtain a surface plasmon wave traveling along that interface (z-axis) in the form [7].

$$E_{x,y} = E_{x,y}^{d,s}\delta_{d,s} \tag{4.1}$$

$$\delta_{d,s} = e^{-i\omega t}e^{i\gamma_{d,s}}e^{i\beta z} \tag{4.2}$$

where ω is the angular frequency, and $\gamma_{d,s}$ and β are the transverse and longitudinal propagation constants, respectively.

This assumed form of an evanescent wave is substituted in the Maxwell's equations:

$$\nabla \times H = J_q + \varepsilon_o \frac{\partial E}{\partial t} + \frac{\partial P}{\partial t} \tag{4.3}$$

$$\nabla \times E = -\mu_o \frac{\partial H}{\partial t} \tag{4.4}$$

$$\nabla . H = 0 \tag{4.5}$$

$$\nabla . \left(\varepsilon_o \frac{\partial E}{\partial t} + \frac{\partial P}{\partial t} \right) = \frac{\partial \rho}{\partial t} \tag{4.6}$$

$$D = \in E \tag{4.7}$$

$$J = \sigma E \tag{4.8}$$

To complete the development of optical plasmon in semiconductors theoretically, the notation of damping has to be described. From a basic perspective, the result of the surface plasmon dispersion equation becomes complex when losses are accounted for. This is directly related to the complex dielectric constant; in order to characterize the dispersion, damping, and excitation of the plasmon, its imaginary part needs to be included. Then the dielectric permittivity equation becomes complex. The dielectric constant is one of the most important factors to assess for future technology applications [6].

Applying the appropriate boundary conditions on both sides of the interface yields the dispersion equation below as:

$$\left(\varepsilon_S^2 - \varepsilon_D^2\right)\left[\frac{k^2\varepsilon_S\varepsilon_D - \beta^2\varepsilon_S - \beta^2 \in_D}{\varepsilon_S\varepsilon_D(\varepsilon_S + \varepsilon_D)}\right] = 0 \tag{4.9}$$

where,

$$k = \frac{\omega}{c} \tag{4.10}$$

ε_S is the dielectric permittivity of the semiconductor, ε_D is the dielectric permittivity of the dielectric, c is the speed of light, and k is the wavenumber.

To be able to account for losses, we used a complex dielectric constant in the form:

$$\varepsilon_r = \varepsilon' + j\varepsilon'' \tag{4.11}$$

where, ε_r is the relative permittivity, $\varepsilon^{'}$ is a real part, and $\varepsilon^{''}$ is the imaginary part. ε_r can be represented by different models, so before presenting our proposed Drude-Lorentz model, we present the Drude model, which is commonly used for metals and highly doped semiconductors.

4.2.1 Drude Model

This model was proposed by Paul Drude in 1900 to explain the transport properties of electrons in metals [9, 10] and has also been adapted for semiconductors [11, 12]. The Drude dielectric function is given by

$$\varepsilon_r = \varepsilon_\infty - \frac{\omega_p^2}{\omega(\omega + j\gamma)} \tag{4.12}$$

where ε_∞ is the high frequency permittivity,γ is the damping term, and ω_p is the plasma angular frequency given by [13]

$$\omega_p = \sqrt{\frac{ne^2}{\varepsilon_o m*}} \tag{4.13}$$

$m*$, e, and n are the electron effective mass, the electron charge, and the carrier density, respectively, and ε_o is the permittivity of free space.

The Drude model is the simplest classical treatment of optical properties of metals. It considers the valence electrons of the atoms to be free. In addition, it is used for semiconductors when free carrier density introduced through doping is sufficiently high to cause the semiconductor to behave similarly to a simple metal. However, in reality, this model has limitations. For example, it does not account for spatial dispersion, which exists when the dielectric constant depends on the wave-vector. Moreover, it does not account for the bound electrons and holes in semiconductors. On the other hand, it does if one replaces ε_o with $\varepsilon_o\varepsilon_r$.

Additionally, recent work by Cada [7] shows that, for more general case, when, $\varepsilon_S \neq \varepsilon_D \neq 1$ a new solution can exist only in a semiconductor/dielectric interface. This may be clearly seen in Eqs. (4.5) and (4.6) in [7] which explain why it cannot appear in the metal.

4.2.2 Lorentz Model

The Lorentz model can be used to describe the frequency response of many materials and typically shows strong dispersion around the resonant frequency [14–16]. It is mostly suited for materials that have bound electrons, with the possibility of having many oscillators in a given system [9, 13, 17–19]. The expression of the dielectric function for a single Lorentz oscillator is given by:

$$\varepsilon_r = \varepsilon_\infty + \frac{\nabla_\in \omega_p^2}{-\omega^2 + j\gamma\omega + \omega_o} \tag{4.14}$$

where ω_o is the resonance frequency and is considered to be equal to an energy band gap of a semiconductor. Δ_ε is a weighting factor given by, $\Delta_\varepsilon = \varepsilon_{st} - \varepsilon_\infty$, with $\in_{st}$ being the static permittivity.

As mentioned, the Lorentz model usually shows strong dispersion around the resonant frequency [14, 18] and is valid only when the photon energy is well below the band gap of the semiconductor. Thus, it cannot be used alone to describe the permittivity of semiconductors.

To overcome this limitation, we propose the use of the Drude-Lorentz model to describe the dielectric function of semiconductors in the Optical frequency range. The proposed dielectric function is given by:

$$\varepsilon_r = \varepsilon_\infty - \frac{\omega_p^2}{\omega(\omega + j\gamma)} + \frac{\Delta_\in \omega_p^2}{-\omega^2 + j\gamma\omega + \omega_o} \tag{4.15}$$

This model is chosen to enable us to take advantage of semiconductors at optical frequencies and search for a possibility of an optical plasmon existence in that range.

4.3 Discussion and Result

A Matlab symbolic tool has been used to implement the above models. Using Eq. (4.9), we insert the desired model of the dielectric function and proceed to calculate the dispersion relation. Before we examine the possible development of the models that work for both materials (i.e. semiconductors and metals), we need to address the question of why it should even be done. The basic answer is to gauge the effect of the loss of plasmons, especially in semiconductors. The slight change in a dielectric permittivity and the damping values are taken into account.

We commence by discussing the details of all of the above models. To solve this dilemma, it is necessary to assume that the system is lightly damped. Based on this assumption, we can then ignore the damping term at first, and later add the loss term for comparison. Next, we recall Eqs. (4.3)–(4.8) and substitute them in the model equation. Doing so, and satisfying the boundary conditions for fields in both media, one can find the dispersion equation for the Drude model without loss as [7]:

$$\begin{aligned} &-\omega^6\left(\varepsilon_{\mathrm{s}} \in_{\mathrm{d}}^2 - \varepsilon_{\mathrm{s}}^2 \in_{\mathrm{d}}\right) + \omega^2\left(\beta^2 c^2 \varepsilon_s^2 - \beta^2 c^2 \varepsilon_{\mathrm{d}}^2 + \varepsilon_{\mathrm{d}}^2 \omega_{\mathrm{p}}^2 - 2\varepsilon_{\mathrm{s}}\varepsilon_{\mathrm{d}}\omega_{\mathrm{p}}^2\right) \\ &\quad + \omega^2\left(2\varepsilon_{\mathrm{d}}\beta^2 c^2 \omega_{\mathrm{p}}^2 + \varepsilon_{\mathrm{d}}\omega_{\mathrm{p}}^2\right) - \beta^2 c^2 \omega_{\mathrm{p}}^2 \\ &\quad = 0 \end{aligned} \tag{4.16}$$

where, ε_s is a semiconductor dielectric permittivity, ε_D is a dielectric permittivity, and ω_p is the plasma frequency. It may be noted that neglecting the losses in this

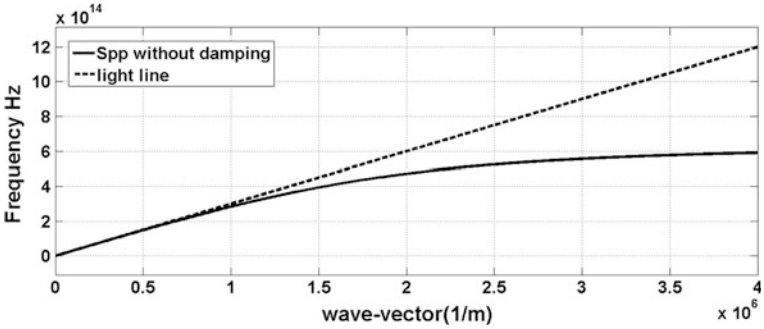


Fig. 4.1 Plasmon dispersion without losses using a Drude model

equation and setting $\varepsilon_S = \varepsilon_D = 1$ [7] for a classical metal/air interface yields the well-known Plasma dispersion equation with well-known solution:

$$-\omega_p^2\omega^4 - \omega^2\left(2\beta^2c^2\omega_p^2 + \omega_p^4\right) + \beta^2c^2\omega_p^4 = 0 \qquad (4.17)$$

Figure 4.1 shows the dispersion relation of surface plasmons propagating along a silicon/air boundary

Employing the generalized Drude theory, the complex dispersion equation is obtained. Equation (4.18) below is the resulting Plasma dispersion equation obtained when losses are included in the Drude model, through the damping frequency, γ.

$$\begin{aligned}&-\left(\omega^2\left(3c^2\omega_p^4\beta^2 - 2c^2\omega_p^2\beta^2\gamma^2 + \omega_p^6\right)\right.\\&\left.-\omega^4\left(2c^2\omega_p^2\beta^2 + 2\omega_p^4 - \omega_p^2\gamma^2\right) + \omega_p^2\omega^6 - c^2\omega_p^6\beta^2 + c^2\omega_p^2\beta^2\gamma^2\right) = 0\end{aligned} \qquad (4.18)$$

For large values of the propagation constant, the surface plasmon frequency approaches a constant value, which can be obtained from Eq. (4.18) and is given by

$$\omega_\infty = \left(\sqrt{(9\omega p^4 - 12\omega p^2\gamma^2 - 8\omega p^2 + 4\gamma^4 + 8\gamma^2)/4} + \left(3\omega p^2\right)4 - \gamma^2/212\right) \qquad (4.19)$$

As can be seen from the dispersion relationship plotted in Fig. 4.2, when damping is taken into account, we observe a drop of the plasmon dispersion from 590 THz to about 432 THz for $\gamma > \omega_p$. However, when $\omega_p > \gamma$ the Eq. (4.19) yields an ω_∞ value of 743 THz which is higher than what we obtain from Fig. 4.1. It is interesting to note that these plasmons are in the Optical frequency range.

Following the same steps as previously and by inserting the Drude-Lorentz model function in Eq. (4.9) and rearranging the terms, one obtains:

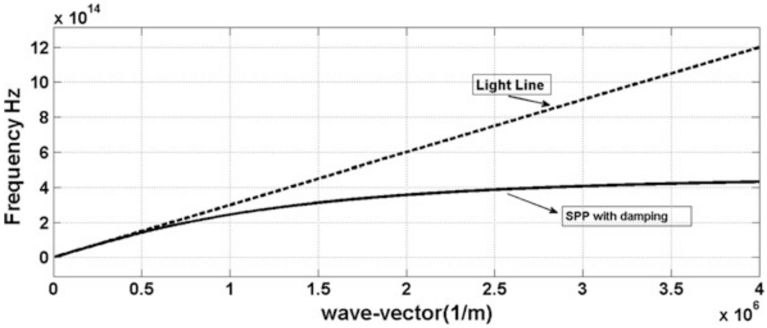


Fig. 4.2 Plasmon dispersion with damping

$$
\begin{aligned}
&-(\omega^2\left(\varepsilon_s^2\omega_p^2+\beta^2c^2\Delta_\varepsilon\omega_p^2+\varepsilon_s\Delta_\varepsilon^2\omega_p^4-2\varepsilon_s\Delta_\in\omega_p^4+\varepsilon_d\omega_p^4\right)+\omega^6\left(\varepsilon_s\varepsilon_d^2-\varepsilon_s^2\varepsilon_d\right)\\
&+\omega^4\left(\varepsilon_s^2\omega_p^2+\beta^2c^2\varepsilon_s^2\omega_p^2-\beta^2c^2\varepsilon_d^2-2\varepsilon_s\varepsilon_d\omega_p^2-\varepsilon_s^2\Delta_\in\omega_p^2+2\Delta_\varepsilon\varepsilon_s\varepsilon_d\omega_p^2\right)\beta^2c^2\omega_p^4\\
&+2\beta^2c^2\Delta_\varepsilon\omega_p^4-\beta^2c^2\Delta_\varepsilon^2\omega_p^4)=0
\end{aligned}
\tag{4.20}
$$

When losses are included in the Drude-Lorentz model, the following dispersion relation is obtained:

$$
\begin{aligned}
&-\omega^4\left(2c^2\omega_p^2\beta^2\gamma^2+2\omega_p^4\gamma^2-\omega_p^4\gamma^2\right)-\omega^2\left(3c^2\omega_p^4\beta^2\gamma^2+2\beta^2\gamma^4\omega_p^4c^2+\omega_p^6\gamma^2\right)\\
&+\gamma^2\omega_p^2\omega^6-c^2\omega_p^4\beta^2\gamma^2+c^2\omega_p^6\beta^2\gamma^2=0
\end{aligned}
\tag{4.21}
$$

This dispersion relation of surface plasmon waves along silicon/air boundary including losses is shown in Fig. 4.3. n-doped silicon is used in the calculations, and all physical parameters are experimental values taken from [20]. To obtain the plasma frequency in the optical range, we had to use high carrier concentrations of the order of 10^{21}/cm^3, which is feasible in silicon. By varying the plasma frequency parameter ω_p of the semiconductor, we obtained a different curve, which has a different surface plasmon frequency, as graphed in Fig. 4.3.

Figure 4.3 shows a different SPP curve than Fig. 4.2, depicting only the real part of the Drude-Lorentz model. When $\omega_p > \gamma$, the surface plasmon frequency increases to approximately 723 THz,. It should be noted that the values of ω_p and γ for n-doped silicon change with doping concentration.

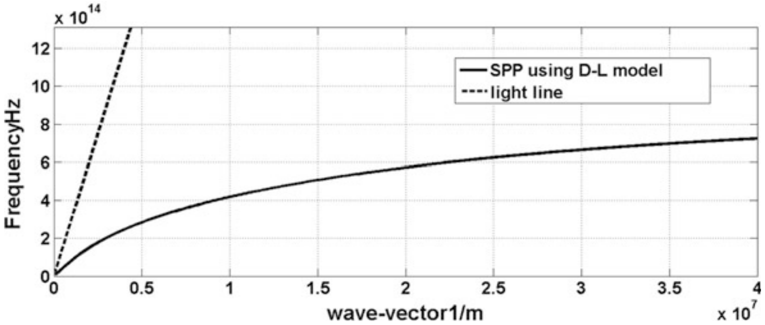


Fig. 4.3 The model surface Plasma dispersion with damping—Drude-Lorentz

4.4 Conclusion and Future Work

Theoretical and numerical studies were conducted on plasmonic interactions at a semiconductor/dielectric interface and a brief review of the basic model theory was presented. We have shown that the inclusion of losses reduces the surface plasmon frequency. As well, we have proposed the Drude-Lorentz model as a model for the dielectric function of semiconductors due to its ability to describe both free electron and bound systems simultaneously. The numerical results of the plasmon dispersion for a silicon/air interface were presented using the proposed model and were compared to the Drude model.

Future work involves the use of other types of semiconductors such as AlGaAs, InP or InGaAs with different and flexible optical properties Different dielectric materials can also be used to tune the surface-plasmon frequency.

Acknowledgments This work was supported by NSERC (Natural Sciences and Engineering Research Council) and NSERC'S ASPIRE (Applied Science in Photonics and Innovative Research in Engineering), both of Canada. We gratefully acknowledge helpful discussions with members of professor Cada's photonics research, and with Professor J. Pištora of technical University of Ostrava, Czech Republic.

References

1. Ritchie RH (1957) Plasma losses by fast electrons in thin films. Phys Rev 106:874
2. Raether H (1988) Surface plasmons on smooth and rough surfaces and on gratings. Springer-Verlag, Berlin, pp 15–22
3. Maier S (2007) Plasmonics fundamentals and applications, 1st edn. Springer Science and business Media LLC, Berlin, pp 21–39
4. Novotny L, Hecht B (2008) Principles of nano-optics, 2nd edn. Cambridge University Press, Cambridge, MA, pp 378–414
5. Zayatsa AV, Smolyaninovb II, Maradudinc AA (2005) Nano-optics of surface plasmon polaritons. Phys Rep 408:131–314

6. Yao B, Fang ZB, Zhu YY, Ji T, He G (2012) A model for the frequency dispersion of the high-k metal-oxide semiconductor capacitance in accumulation. Appl Phys Lett 100: 222903/1–3
7. Cada M, Pištora J (2011) Optical plasmons in semiconductors. In: ISMOT conference, June 20–23
8. Eldlio M, Che F, Cada M (2012) Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science, WCECS 2012, 24–6 Oct, 2012, San Francisco, USA, pp 1078–1081
9. Fox M (2010) Optical properties of solids, 2nd edn. Oxford University Press, Oxford, pp 33–44
10. Chuang S (2009) Physics of photonic devices, 2nd edn. Wiley, New York, pp 193–196
11. Kuttge M, Kurz H, Rivas JG, Sánchez-Gil JA, Bolívar PH (2007) Analysis of the propagation of terahertz surface plasmon polaritons on semiconductor groove gratings. J Appl Phys 101(2):023707-1–023707-6
12. West PR, Ishii S, Naik GV, Emani NK, Shalaev VM, Boltasseva A (2010) Searching for better plasmonic materials. Laser Photonics Rev 4(6):795–808
13. Lee KH, Ahmed I, Goh RSM, Khoo EH, Li EP, Hung TGG (2011) Implementation of the FDTD method based on the Lorentz-Drude model on GPU for plasmonics applications. Prog Electromagnet Res 116:441–456
14. Rivas JG, Snchez-Gil JA, Kuttge M, Bolivar PH, Kurz H (2006) Optically switchable mirrors for surface plasmon polaritons propagating on semiconductor surfaces. Phys Rev B 74:245–324
15. Huang Y, Ho ST (2006) Computational model of solid state, molecular, or atomic media for FDTD simulation based on a multilevel multi-electron system governed by Pauli exclusion and Fermi-Dirac thermalization with application to semiconductor photonics. Opt Express 14:3569–3587
16. Hryciw A, Jun YC, Brongersma ML (2010) Plasmonic: electrifying plasmonics on silicon. Nat Mater 9:3–4
17. Janke C, Rivas JG, Bolivar PH, Kurz H (2005) All optical switching of electromagnetic radiation through subwavelength apertures. Opt Lett 30(18):2357–2359
18. Ahmed I, Khoo E, Kurniawan O, Li E (2011) Modeling and simulation of active plasmonics with the FDTD method by using solid state and Lorentz–Drude dispersive model. J Opt Soc Am B 28(3):325–359
19. van Exter M, Grischkowsky D (1990) Optical and electronics properties of doped silicon from 0.1 to 2 THz. Appl Phys let 56(17):1694–1696
20. Laghla Y, Scheid E (1997) Optical study of undoped, B or P-doped polysilicon. Thin Solid Film 306:67–73

Chapter 5
Optimal Power Conversion of Standalone Wind Energy Conversion Systems Using Fuzzy Adaptive Control

Hoa M. Nguyen and D. Subbaram Naidu

Abstract This chapter presents an advanced control technique to deal with the maximum power conversion problem of a standalone Wind Energy Conversion Systems (WECS) with Permanent Magnet Synchronous Generators (PMSG). The proposed method, which is different from traditional Maximum Power Point Tracking (MPPT) methods, is based on a fuzzy adaptive control scheme in which the adaptation is obtained from the *Lyapunov* analysis and carried out by the fuzzy logic technique. The superiority of the advanced control technique is shown by numerical simulations with comparison between the proposed controller's performance and a nonlinear feedback linearization controller's performance.

Keywords Fuzzy adaptive control · *Lyapunov* analysis · Nonlinear feedback linearization control · Optimal power conversion · Permanent magnet synchronous generator · Standalone wind energy conversion system.

5.1 Introduction

Wind energy is an attractive energy source due to its clean, renewable, and abundant features and has become the fastest growing source among other renewable energy sources in the last few decades [1–3]. Wind is converted into energy basically by wind turbines or Wind Energy Conversion Systems (WECS). These wind energy converters can be classified as grid-connected or standalone

H. M. Nguyen · D. S. Naidu (✉)
Department of Electrical Engineering, Idaho State University,
921 S. 8th Avenue Pocatello, ID 83209, USA
e-mail: naiduds@isu.edu

H. M. Nguyen
e-mail: nguyhoa@isu.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_5,

depending on their connection status to utility grids or local grids. Nowadays most WECS are connected to utility grids. However, there is still demand for standalone WECS which provide electrical power to remote areas where utility grids are not available. To guarantee continuous energy supply, standalone WECS are combined with other energy sources such as battery storage systems, solar energy systems, diesel generators, etc., resulting in Hybrid Wind Energy Systems (HWES). Due to the presence of other energy sources, there are two primary control objectives in the HWES. One is to maximize the power conversion of the standalone WECS under stochastic wind changes, and the other is to ensure constant power flow to local loads. This chapter is focused on the first control objective of maximum power conversion. Works related to the second control objective are available from [4–7].

The optimal power conversion problem in WECS is traditionally solved by Maximum Power Point Tracking (MPPT) methods [8, 9] where controllers are designed to find and maintain maximum power points as the wind speed changes. This control scheme is quite simple because it does not require dynamical models of WECS in most cases. However, the main drawback of the MPPT methods is the tradeoff between the power tracking and control efficiency. As a result, more advanced control techniques for improving the problem become important.

Various advanced control approaches have been proposed to deal with the maximum power conversion of standalone WECS such as the nonlinear feedback linearization [10], sliding mode control [11], and adaptive control [12]. Unlike the approach proposed in [12] where the learning rule is realized by neural networks, this chapter proposed an adaptive scheme based on the input-output linearization control where the learning rule is implemented by a fuzzy logic technique. To show the advantage of the adaptive method over the nonlinear feedback linearization method, the standalone WECS model is chosen as the same as in [10] and the performance of the two controllers are compared via simulations. It should be emphasized that this chapter is an extended and revised version of the work presented in [13].

The chapter is organized as follows. Section 5.2 describes the HWES and the standalone PMSG-based WECS nonlinear model. The adaptive control design is presented in Sect. 5.3 followed by its application to the standalone WECS in Sect. 5.4. Based on simulation results shown in Sect. 5.5, some discussions and conclusions are drawn in Sect. 5.6.

5.2 Standalone Wind Energy Conversion System Modeling

A HWES includes a WECS interacting with another source of energy as shown in Fig. 5.1. Due to the output power from the WECS fluctuating according to wind changes, other energy sources such as battery or solar systems or diesel generators must be added to ensure a constant power supply to the local grid. Maximum power conversion of the WECS is obtained by adjusting the generator speed, ω_g, as

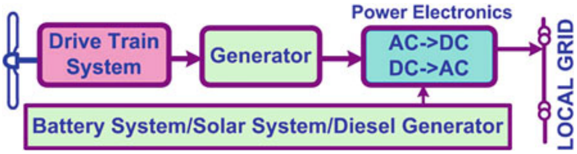


Fig. 5.1 Hybrid wind energy system (HWES)

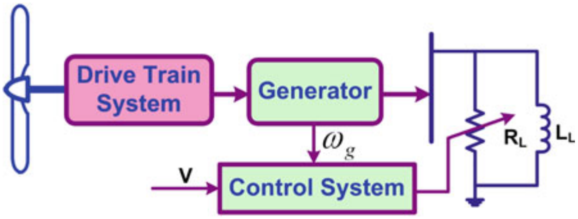


Fig. 5.2 Standalone WECS

wind speed, V, changes. This is achieved by modifying the equivalent load at the generator terminal via power electronics converters. The equivalent standalone WECS is depicted in Fig. 5.2 where R_L and L_L are the equivalent load resistance and inductance, respectively. The equivalent load resistance is considered as the control input for the control system.

The dynamic model of the standalone WECS is obtained by combining the aerodynamics, drive train dynamics and generator dynamics. Note that the power electronics dynamics is ignored because it is much faster than the other dynamics.

The aerodynamics converts wind flows into aerodynamic torque and mechanical power given respectively as

$$T_r = \frac{1}{2}\rho\pi R^3 V^2 C_Q(\lambda), \tag{5.1}$$

$$P_r = \frac{1}{2}\rho\pi R^2 V^3 C_P(\lambda), \tag{5.2}$$

where T_r is the aerodynamic torque, ρ is the air density, R is the radius of the wind rotor swept area, V is the wind speed, $C_Q(\lambda)$ is the torque coefficient, P_r is the mechanical power, and $C_P(\lambda)$ is the power coefficient. It is seen that both torque and power coefficient are functions of the so-called tip-speed ratio λ which characterizes the power conversion efficiency of the wind turbine. The tip-speed ratio is defined as the ratio between the speed at the tip of blades and the wind speed, which is given as

$$\lambda = \frac{\omega_r R}{V}, \tag{5.3}$$

where ω_r is the wind rotor rotational speed. The torque coefficient in (5.1) can be approximated as the following sixth-order polynomial function of the tip-speed ratio [10]

$$C_Q(\lambda) = a_6\lambda^6 + a_5\lambda^5 + a_4\lambda^4 + a_3\lambda^3 + a_2\lambda^2 + a_1\lambda + a_0. \tag{5.4}$$

It is well known that the relationship between the power and torque coefficients is given by

$$C_Q(\lambda) = \frac{C_P(\lambda)}{\lambda}. \tag{5.5}$$

The power coefficient $C_P(\lambda)$ has its maximum value at the so-called optimal tip-speed ratio λ^* as illustrated in Fig. 5.3. Therefore, to maximize the power conversion, the WECS must operate at the optimal tip-speed ratio. However, when the wind speed changes, the tip-speed ratio is perturbed away from the optimal value as seen from (5.3). Therefore, to maintain the optimal tip-speed ratio, the wind rotor speed ω_r must be adjusted by the control system.

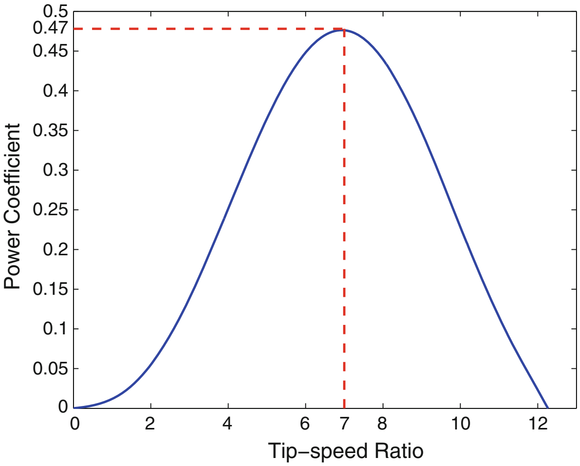
The standalone PMSG model in the *direct* and *quadrature* (d, q) axes/frame is given as [10]

$$\frac{d}{dt}i_d = -\frac{R_s + R_L}{L_d + L_L}i_d + \frac{p(L_q - L_L)}{L_d + L_L}i_q\omega_g, \tag{5.6}$$

$$\frac{d}{dt}i_q = -\frac{R_s + R_L}{L_q + L_L}i_q - \frac{p(L_d + L_L)}{L_q + L_L}i_d\omega_g + \frac{p\Phi_m}{L_q + L_L}\omega_g, \tag{5.7}$$

$$T_g = p\Phi_m i_q, \tag{5.8}$$

Fig. 5.3 Power coefficient curve versus tip-speed ratio



where i_d and i_q are the d- and q- components of the stator currents respectively; L_d and L_q are the d- and q-components of the stator inductances respectively; R_s is the stator resistance; R_L is the equivalent load resistance; p is the number of pole pairs; Φ_m is the linkage flux; ω_g is the high-speed or generator speed; and T_g is the generator electromagnetic torque.

The drive train system consists of a low-speed shaft connected to a high-speed shaft through a gearbox which is a rotational speed multiplier. The drive train dynamics can be represented by a rigid model as

$$J_h \frac{d\omega_g}{dt} = \frac{\eta}{i} T_r - T_g, \tag{5.9}$$

where J_h is the equivalent inertia transformed into the high-speed side, η and i are the gearbox efficiency and speed ratio, respectively, T_r is the aerodynamic torque, and T_g is the generator electromagnetic torque.

Define $\mathbf{x} = [x_1\, x_2\, x_3]^T = [i_d\, i_q\, \omega_g]^T$ as the state variable vector, $u = R_L$ as the control input, and $y = \omega_g$ as the system output. Combining (5.1), (5.3), and (5.6)–(5.9) results in a complete nonlinear state space model of the standalone PMSG-based WECS:

$$\underbrace{\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix}}_{\dot{x}} = \underbrace{\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \end{bmatrix}}_{f(x)} + \underbrace{\begin{bmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ g_3(\mathbf{x}) \end{bmatrix}}_{g(x)} u, \tag{5.10}$$

$$y = h(\mathbf{x}), \tag{5.11}$$

where

$$f_1(\mathbf{x}) = -\frac{R_s}{L_d + L_L} x_1 + \frac{p(L_q - L_L)}{L_d + L_L} x_2 x_3, \tag{5.12}$$

$$\begin{aligned} f_2(\mathbf{x}) = &-\frac{R_s}{L_q + L_L} x_2 - \frac{p(L_d + L_L)}{L_q + L_L} x_1 x_3 \\ &+ \frac{p\Phi_m}{L_q + L_L} x_3, \end{aligned} \tag{5.13}$$

$$f_3(\mathbf{x}) = \frac{\eta\rho\pi R^3 V^2}{2iJ_h} C_Q(x_3, V) - \frac{p\Phi_m}{J_h} x_2, \tag{5.14}$$

and

$$g_1(\mathbf{x}) = -\frac{1}{L_d + L_L} x_1, \tag{5.15}$$

$$g_2(\mathbf{x}) = -\frac{1}{L_q + L_L} x_2, \tag{5.16}$$

$$g_3(\mathbf{x}) = 0, \tag{5.17}$$

$$h(\mathbf{x}) = x_3. \tag{5.18}$$

Note that the wind rotor speed ω_r is multiplied i times after going through the gearbox, therefore the generator speed ω_g is i times larger than the wind rotor speed. Consequently the torque coefficient $C_Q(x_3, V)$ in (5.14) can be expressed as

$$\begin{aligned} C_Q(x_3, V) = a_6\left(\frac{Rx_3}{iV}\right)^6 + a_5\left(\frac{Rx_3}{iV}\right)^5 + a_4\left(\frac{Rx_3}{iV}\right)^4 + a_3\left(\frac{Rx_3}{iV}\right)^3 + a_2\left(\frac{Rx_3}{iV}\right)^2 \\ + a_1\left(\frac{Rx_3}{iV}\right) + a_0. \end{aligned} \tag{5.19}$$

It is observed from (5.19) that the dynamical model (5.10)–(5.11) is highly nonlinear.

5.3 Adaptive Control Method

Consider a Single Input Single Output (SISO) nonlinear system defined in the region $D_x \in R^n$ as

$$\dot{\mathbf{x}} = f(\mathbf{x}) + g(\mathbf{x})u, \tag{5.20}$$

$$y = h(\mathbf{x}), \tag{5.21}$$

where $\mathbf{x} \in R^n$ is the state vector, $u \in R^1$ is the control input, $y \in R^1$ is the system output, $f(\mathbf{x}) \in R^n$ and $g(\mathbf{x}) \in R^n$ are smooth vector fields, and $h(\mathbf{x}) \in R^1$ is a scalar smooth function. It is assumed the nonlinear system (5.20)–(5.21) has a relative degree r $(r < n)$ at $x_0 \in D_x$ and internal dynamics is stable. Taking derivatives of the output y with respect to time up to r times gives

$$y^{(r)} = L_f^r h(\mathbf{x}) + \underbrace{L_g L_f^{r-1} h(\mathbf{x})}_{\neq 0} u, \tag{5.22}$$

where $L_f^r h(\mathbf{x})$ is the *Lie* derivative of $h(\mathbf{x})$ along the direction of the vector field $f(\mathbf{x})$ up to r times, $L_g L_f^{r-1} h(\mathbf{x})$ is the *Lie* derivative of $L_f^{r-1} h(\mathbf{x})$ a long the direction of the vector field $g(\mathbf{x})$.

Defining $\alpha(\mathbf{x}) = L_f^r h(\mathbf{x})$ and $\beta(\mathbf{x}) = L_g L_f^{r-1} h(\mathbf{x})$ and rewriting (5.22) yields

$$y^{(r)} = \alpha(\mathbf{x}) + \beta(\mathbf{x})u. \tag{5.23}$$

Assumption 1 The function $\beta(\mathbf{x})$ is positive (i.e., $0 < \beta(\mathbf{x}) < \infty$ with $\forall x \in D_x$) [14, 15].

Choosing the state feedback linearization control:

$$u^*(\mathbf{x}) = \frac{1}{\beta(\mathbf{x})}(-\alpha(\mathbf{x}) + v), \tag{5.24}$$

then the input-output nonlinear relation (5.23) becomes linear as

$$y^{(r)} = v. \tag{5.25}$$

5.3.1 Fuzzy Approximation Strategy

The linear input-output relation (5.25) can be designed for stabilization or tracking with any linear methods. It is apparent that the ideal feedback linearization control (5.24) is only applicable if $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are known. However, there are uncertainties imposed on $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ in practice, therefore the control (24) is not accurate. In this case, adaptive algorithms were proposed to realize the ideal control $u^*(\mathbf{x})$ by using an approximate nonlinear function $\hat{u}(\mathbf{x})$, called direct adaptive control, or by using approximate $\hat{\alpha}(\mathbf{x})$ and $\hat{\beta}(\mathbf{x})$ of nonlinear functions $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$, called indirect adaptive control [14–16]. Here the emphasis is placed on the direct adaptive control using the fuzzy approximation.

The control problem is to drive the output y track a reference signal y_m (y_m is a smooth function). The feedback linearized control input v in (5.25) is defined as

$$v = y_m^{(r)} + \bar{e}_s + \gamma e_s, \tag{5.26}$$

where γ is a positive constant, $\bar{e}_s$ and e_s are defined as

$$e_s = e_o^{(r-1)} + k_1 e_o^{(r-2)} + \ldots + k_{r-1} e_o, \tag{5.27}$$

$$\bar{e}_s = \dot{e}_s - e_o^{(r)} = k_1 e_o^{(r-1)} + \ldots + k_{r-1}\dot{e}_o, \tag{5.28}$$

$$e_o = y_m - y, \tag{5.29}$$

where e_s is the tracking error, e_o is the output error. Coefficients k_i are chosen such that the following polynomial is *Hurwitz*

$$E(s) = s^{r-1} + k_1 s^{r-2} + \ldots + k_{r-2}s + k_{r-1}. \tag{5.30}$$

The ideal control input in (5.24) is approximated by a fuzzy system as

$$\hat{u}(\mathbf{x}) = \theta_u^T \xi_u(\mathbf{x}), \tag{5.31}$$

where θ_u is a parameter vector including values of singleton membership function at consequent propositions of the fuzzy system rule base, $\xi_u(\mathbf{x})$ is a fuzzy regressive vector. θ_u is updated online such that $\hat{u}(\mathbf{x})$ approaches $u^*(\mathbf{x})$. The optimal parameter vector is

$$\theta_u^* = \arg \min_{\theta \in D_\theta} \left\{ \sup_{x \in D_x} \left| \theta_u^T \xi_u(\boldsymbol{x}) - u^* \right| \right\}. \tag{5.32}$$

Because $u^*(\mathbf{x})$ is approximated by a fuzzy system possessing a finite number of rules, there exists an unavoidable structural error $\delta_u(\mathbf{x})$. Therefore the actual ideal control $u^*(\mathbf{x})$ is

$$u^*(\mathbf{x}) = \theta_u^{*T} \xi_u(\mathbf{x}) + \delta_u(\mathbf{x}). \tag{5.33}$$

The difference between the approximate control $\hat{u}(\mathbf{x})$ and ideal control $u^*(\mathbf{x})$ is

$$\hat{u}(\mathbf{x}) - u^*(\mathbf{x}) = \tilde{\theta}_u^T \xi_u(\mathbf{x}) - \delta_u(\mathbf{x}), \tag{5.34}$$

where

$$\tilde{\theta}_u = \theta_u - \theta_u^* \tag{5.35}$$

is the approximation error.

Assumption 2 The adaptive control is chosen such that the structural error is bounded $\left(\left|\delta_u(\mathbf{x})\right| \le \bar{\delta}_u\right)$ with $\forall x \in D_x$ and the upper bound $\bar{\delta}_u$ is known [14, 15].

Due to the presence of the structural error, an additional supervisory control u_s is added to guarantee the closed-loop stability. Therefore the revised control is

$$u = \hat{u} + u_s. \tag{5.36}$$

5.3.2 Adaptive Law Design

The adaptive law, designed based on *Lyapunov* analysis, is presented here.

Adding and subtracting $\beta(\mathbf{x})u^*(\mathbf{x})$ into (5.23) gives

$$\begin{aligned} y^{(r)} &= \alpha(\mathbf{x}) + \beta(\mathbf{x})u^*(\mathbf{x}) + \beta(\mathbf{x})[u(\mathbf{x}) - u^*(\mathbf{x})], \\ &= v + \beta(\mathbf{x})[u(\mathbf{x}) - u^*(\mathbf{x})]. \end{aligned} \tag{5.37}$$

Combining (5.26), (5.29), (5.34), (5.36), and (5.37) gives

$$\begin{aligned} e_o^{(r)} &= y_m^{(r)} - y^{(r)}, \\ e_o^{(r)} &= y_m^{(r)} - v - \beta(\mathbf{x})[u(\mathbf{x}) - u^*(\mathbf{x})], \\ e_o^{(r)} &= -\bar{e}_s - \gamma e_s - \beta(\mathbf{x})[\hat{u}(\mathbf{x}) + u_s - u^*(\mathbf{x})], \\ e_o^{(r)} &= -\bar{e}_s - \gamma e_s - \beta(\mathbf{x})\tilde{\theta}_u^T \xi_u(\mathbf{x}) + \beta(\mathbf{x})\delta_u(\mathbf{x}) - \beta(\mathbf{x})u_s. \end{aligned} \tag{5.38}$$

Combining (5.28) and (5.38) yields the error dynamics as

$$\dot{e}_s + \gamma e_s = -\beta(\mathbf{x})\tilde{\theta}_u^T \xi_u(\mathbf{x}) + \beta(\mathbf{x})\delta_u(\mathbf{x}) - \beta(\mathbf{x})u_s. \tag{5.39}$$

Considering a positive semidefinite quadratic *Lyapunov* function

$$V = \frac{1}{2\beta(\mathbf{x})} e_s^2 + \frac{1}{2}\tilde{\theta}_u^T Q_u \tilde{\theta}_u, \tag{5.40}$$

where Q_u is a positive definite weighting matrix. Taking the derivative of V with respect to time, with the observation from (5.35) that $\dot{\tilde{\theta}}_u = \dot{\theta}_u$, yields

$$\dot{V} = \frac{1}{\beta(\mathbf{x})} e_s \dot{e}_s - \frac{\dot{\beta}(\mathbf{x})}{2\beta^2(\mathbf{x})} e_s^2 + \tilde{\theta}_u^T Q_u \dot{\theta}_u. \tag{5.41}$$

Substituting (5.39) into (5.41) produces

$$\begin{aligned} \dot{V} &= \frac{e_s}{\beta(\mathbf{x})}\Big[-\gamma e_s - \beta(\mathbf{x})\tilde{\theta}_u^T \xi_u(\mathbf{x}) + \beta(\mathbf{x})\delta_u(\mathbf{x}) - \beta(\mathbf{x})u_s \Big] - \frac{\dot{\beta}(\mathbf{x})}{2\beta^2(\mathbf{x})} e_s^2 + \tilde{\theta}_u^T Q_u \dot{\theta}_u, \\ &= -\frac{\gamma e_s^2}{\beta(\mathbf{x})} - e_s u_s + e_s \delta_u(\mathbf{x}) + \tilde{\theta}_u^T \left(Q_u \dot{\theta}_u - \xi(\mathbf{x}) e_s \right) - \frac{\dot{\beta}(\mathbf{x})}{2\beta^2(\mathbf{x})} e_s^2. \end{aligned} \tag{5.42}$$

Choosing the adaptive law as

$$\dot{\theta}_u = Q_u^{-1} \xi_u(\mathbf{x}) e_s, \tag{5.43}$$

and substituting (5.43) into (5.42) gives

$$\dot{V} = -\frac{\gamma e_s^2}{\beta(\mathbf{x})} - e_s u_s + e_s \delta_u(\mathbf{x}) - \frac{\dot{\beta}(\mathbf{x})}{2\beta^2(\mathbf{x})} e_s^2, \tag{5.44}$$

$$\dot{V} \leq -\frac{\gamma e_s^2}{\beta(\mathbf{x})} - e_s u_s + |e_s| \left(|\delta_u(\mathbf{x})| + \frac{|\dot{\beta}(\mathbf{x})|}{2\beta^2(\mathbf{x})} |e_s| \right). \tag{5.45}$$

Assumption 3 There exist a positive lower bound and upper bound of $\beta(\mathbf{x})$ (i.e. $0 < \underline{\beta} \leq \beta(\mathbf{x}) \leq \bar{\beta}$) [14, 15].

Assumption 4 The time derivative (velocity) of $\beta(\mathbf{x})$ is bounded (i.e. $|\dot{\beta}(\mathbf{x})| \leq \beta_v$) [14, 15].

Combining (5.45) with Assumptions 3 and 4 yields

$$\dot{V} \leq -\frac{\gamma e_s^2}{\bar{\beta}} - e_s u_s + |e_s| \left(\bar{\delta}_u + \frac{\beta_v}{2\underline{\beta}^2} |e_s| \right). \tag{5.46}$$

Choosing the supervisory control u_s as

$$u_s = \left(\bar{\delta}_u + \frac{\beta_v}{2\underline{\beta}^2} |e_s| \right) sgn(e_s), \tag{5.47}$$

and substituting (5.47) into (5.46) gives

$$\dot{V} \leq -\frac{\gamma e_s^2}{\bar{\beta}} \leq 0. \tag{5.48}$$

Note that $e_s sgn(e_s) = |e_s|$. It is seen from (5.40) and (5.48), the positive semidefinite *Lyapunov* function V has its negative semidefinite derivative $\dot{V}$, therefore the closed-loop adaptive system is stable [17].

5.4 Adaptive Control Design for the Standalone WECS

In this section, the adaptive control method presented in Sect. 5.3 is applied to the standalone nonlinear PMSG-based WECS given in (5.10) and (5.11). The control objective is to track the optimal generator speed reference ω_g^* in order to maintain the optimal tip-speed ratio λ^* as the wind speed V changes. Unlike the nonlinear feedback linearization design in [10] where authors simplified the sixth-order polynomial torque coefficient in (5.4) by using an approximate second-order polynomial which captures only the steady-state region, our study uses the sixth-order polynomial torque coefficient which captures all operating regions.

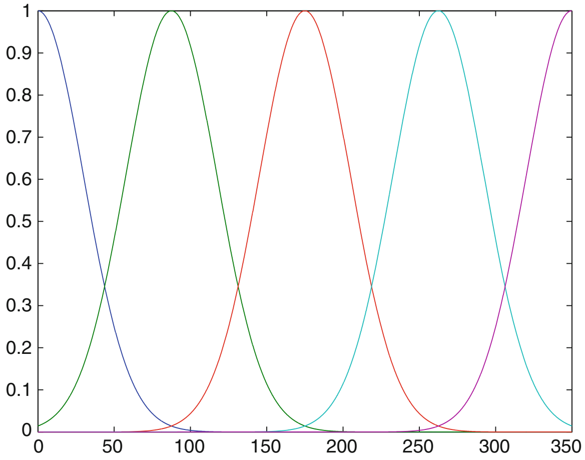


Fig. 5.4 Gaussian fuzzy sets of the linguistic variable x_3

The fuzzy model $\hat{u}(\mathbf{x})$, used to approximate the control $u^*(\mathbf{x})$, is chosen as Takagi-Sugeno model which has inference rules in the form

$$IF\, x_1\, is\, A_{1i}\; and\; x_2\, is\, A_{2i}\; and\; x_3\, is\, A_{3i}\; THEN\, \hat{u}(\mathbf{x}) = \theta_{ui},$$

where x_1, x_2, and x_3 are state variables; A_{1i}, A_{2i}, and A_{3i} are fuzzy sets of linguistic variables describing x_1, x_2, and x_3, respectively at the *ith* rule. The forms and number of fuzzy sets for linguistic variables are chosen based on trial and error basis. In this system fuzzy set forms were chosen as Gaussian and there are five fuzzy sets for each linguistic variable as shown in Fig. 5.4. Consequently there are 125 total rules.

The approximate fuzzy system in (5.31) is

$$\hat{u}(\mathbf{x}) = \theta_u^T \xi_u(\mathbf{x}),$$

where

$$\theta_u = [\theta_{u1}\, \theta_{u2} \ldots \theta_{u125}]^T, \tag{5.49}$$

$$\xi_u(\mathbf{x}) = [\xi_{u1}(\mathbf{x})\; \xi_{u2}(\mathbf{x}) \xi_{u125}(\mathbf{x})]^T, \tag{5.50}$$

$$\xi_{\mathbf{ui}} = \frac{\mu_{A_{1i}}(x_1) \cdot \mu_{A_{2i}}(x_2) \cdot \mu_{A_{3i}}(x_3)}{\sum_{i=1}^{125} \mu_{A_{1i}}(x_1) \cdot \mu_{A_{2i}}(x_2) \cdot \mu_{A_{3i}}(x_3)}, \tag{5.51}$$

where ξ_{ui} is the regressor at the *ith* rule.

The direct fuzzy adaptive control structure is shown in Fig. 5.5 where the adaptive controller is given in (5.31) with the adaptive law given in (5.43), and the supervisory controller is given in (5.47).

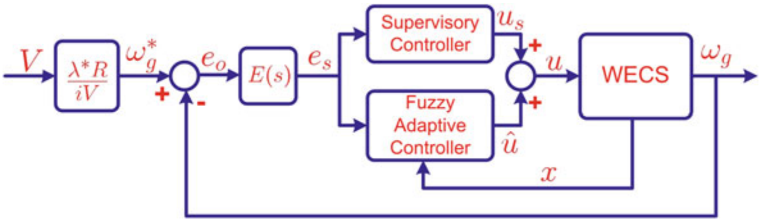


Fig. 5.5 Direct fuzzy adaptive control structure

5.5 Simulation Results

Simulations were carried out with a 3KW standalone PMSG-based WECS which has the optimal power coefficient $C_{P_{max}} = 0.478$ and the optimal tip-speed ratio $\lambda^* = 7$. Other system parameters are given in Table 5.1 [10]. Lower and upper bounds in Assumption 2, 3, and 4 are $\bar{\delta}_u = 0.001$, $\underline{\beta} = 1$, and $\beta_v = 30$. The nonlinear PMSG-based WECS has the relative degree $r = 2$, so parameters for the error dynamics are chosen as $\gamma = 15$ and $k_1 = 5$. The control input is set bounded as $0 < u \leq 100$. The stochastic wind profile is shown in Fig. 5.6. Control performances of both Direct Fuzzy Adaptive Control (DFAC) proposed in this study and Feedback Linearization Control (FLC) proposed in [10] are compared in parallel.

The output tracking performance is shown in Figs. 5.7 and 5.8, where the solid line represents the optimal speed reference computed based on the optimal tip-speed ratio, the dashed-dotted line and dashed line represent the tracking performance of the DFAC and FLC control, respectively. It is difficult to compare the tracking performance of the two controllers by looking at Fig. 5.7; therefore, square tracking errors of the DFAC and FLC are provided in Fig. 5.9 which indicates that the DFAC provides better tracking than the FLC.

To demonstrate the power conversion efficiency of the two controllers under wind speed fluctuations, the power coefficient and tip-speed ratio are shown in Figs. 5.10 and 5.11, respectively. These two figures show that the power coefficient and tip-speed ratio of the wind turbine are maintained at the optimal values, showing that the maximum power conversion efficiency is achieved. In particular, it is obvious from those figures that the DFAC stays constantly steady at the optimal power coefficient and tip-speed ratio values after the transient time. Meanwhile, the FLC keeps oscillating around optimal values. This fact shows the DFAC is better than the FLC with respect to the power conversion efficiency.

Table 5.1 Simulation Data

Parameter	Value	Parameter	Value	Parameter	Value
R	2.5 m	R_s	$3.3\,\Omega$	ρ	$1.25\,kg/m^3$
L_d	$0.04156\,H$	L_q	$0.04156\,H$	L_s	$0.08\,H$
η	1	p	3	J_h	$0.0552\,kg.m^2$
i	7	Φ_m	$0.4382\,Wb$		

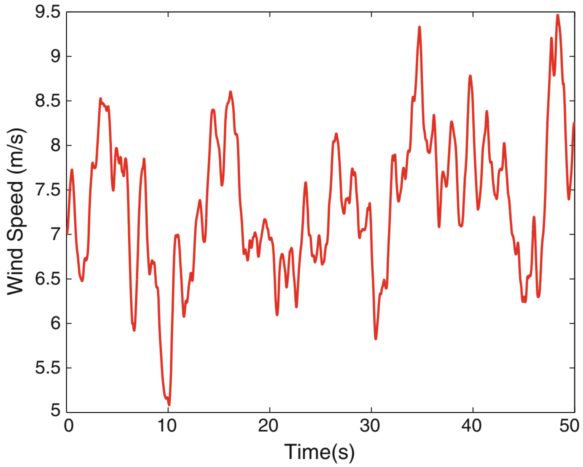


Fig. 5.6 Simulation wind speed profile

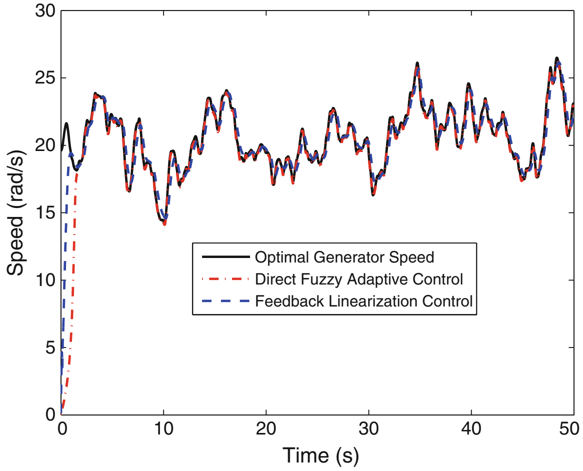


Fig. 5.7 Output reference tracking

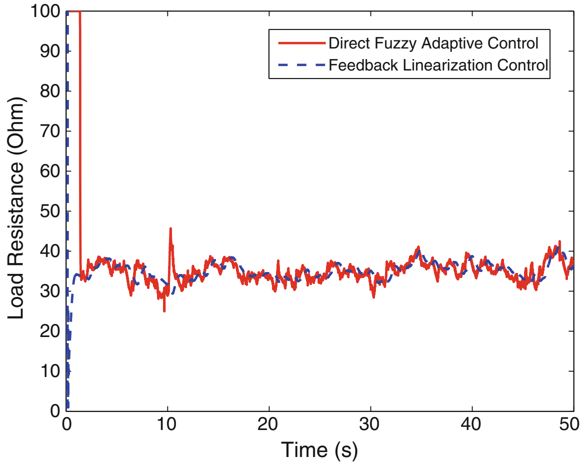


Fig. 5.8 Control inputs

Fig. 5.9 Tracking errors

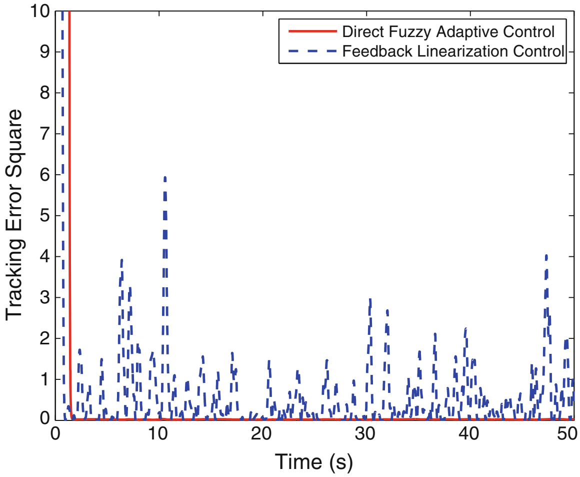
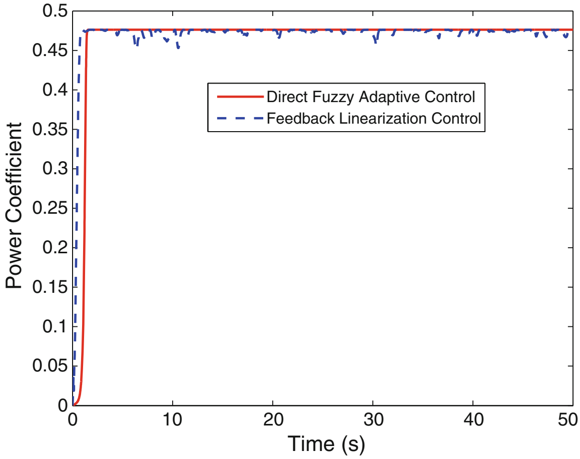


Fig. 5.10 Optimal power coefficient maintenance



5.6 Discussion and Conclusion

The proposed DFAC is better than the FLC in dealing with the time-varying, nonlinear nature of WECS as illustrated by simulation results. The DFAC was also proven more effective than the FLC regarding the control performance and power capture. However, the design of DFAC requires certain assumptions to be met as stated in Sect. 5.3 which are not always satisfied in practice. For example, it is hard to find lower and upper bounds of the nonlinear function $\beta(\boldsymbol{x})$ and structural error $\delta_u(\boldsymbol{x})$. These values were found based on the system physical features combined with the trial and error method in this study. Another important note is that both DFAC and FLC require all states to be accessible. As we know that all system

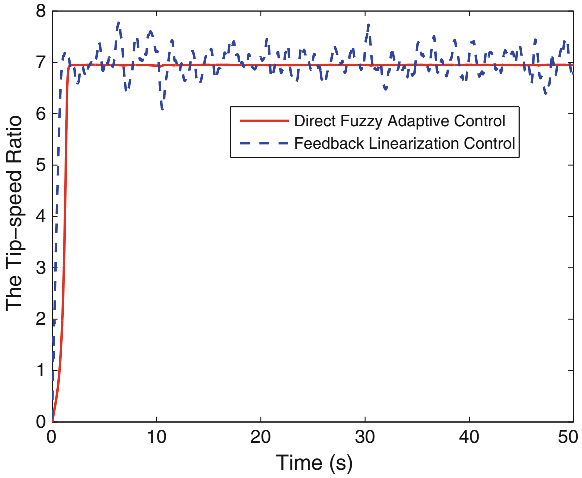


Fig. 5.11 Optimal tip-speed ratio maintenance

states are not available and the DFAC and FLC cannot used directly to address those situations. However, nonlinear state estimators are needed to use the DFAC scheme.

References

1. Musgrove P (2010) Wind power. Cambridge University Press, Cambridge, MA
2. Shepherd W, Zhang L (2011) Electricity generation using wind power. World Scientific Publishing Co.Pte.Ltd, Singapore
3. Stiebler M (208) Wind energy systems for electric power generation. Springer, New York.
4. Abbey C, Strunz K, Joos G (2009) A knowledge-based approach for control of two-level energy storage for wind energy systems. IEEE Trans Energy Convers 24(2):539–547
5. Teleke S, Baran M, Huang A, Bhattacharya S, Anderson L (2009) Control strategies for battery energy storage for wind farm dispatching. IEEE Trans Energy Convers 24(3): 725–732
6. Teleke S, Baran M, Bhattacharya S, Huang A (2010) Optimal control of battery energy storage for wind farm dispatching. IEEE Trans Energy Convers 25(3):787–794
7. Valenciaga F, Puleston P (2005) Supervisor control for a stand-alone hybrid generation system using wind and photovoltaic energy. IEEE Trans Energy Convers 20(2):398–405
8. Agarwal V, Aggarwal R, Patidar P, Patki C (2010) A novel scheme for rapid tracking of maximum power point in wind energy generation systems. IEEE Trans Energy Convers 25(1):228–236
9. Kazmi S, Goto H, Guo H, Ichinokura O (2011) A novel algorithm for fast and efficient speed-sensorless maximum power point tracking in wind energy conversion systems. IEEE Trans Ind Electron 58(1):29–36
10. Munteanu I, Bratcu A, Cutuluslis N, Ceanga E (2008) Optimal control of wind energy systems: toward a global approach. Springer, New York
11. Valenciaga F, Puleston P (2008) High-order sliding control for a wind energy conversion system based on a permanent magnet synchronous generator. IEEE Trans Energy Convers 23(3):860–867

12. Mayosky M, Cancelo I (1999) Direct adaptive control of wind energy conversion systems using Gaussian networks. IEEE Trans Neural Networks 10(4):898–906
13. Nguyen H, Naidu D (2012) Direct fuzzy adaptive control for standalone wind energy conversion systems. In: Lecture notes in engineering and computer science, Proceedings of the world Congress on engineering and computer Science, WCECS 2012, October 24–26, 2012, San Francisco, USA, vol II, pp 994–999 (2012).
14. Huynh HT (2006) Intelligent control systems. Ho Chi Minh National University Press.
15. Spooner J, Passino K (1996) Stable adaptive control using fuzzy systems and neural networks. IEEE Trans Fuzzy Syst 4(3):339–359 16.
16. Wang LX (1996) A course in fuzzy systems and control. Prentice-Hall. Inc., New Jersey
17. Khalil H (1996) Nonlinear systems, 2nd edn. Prentice-Hall. Inc., New Jersey

Chapter 6
Design of Augmented Observer for Rotor Systems

Zhentao Wang, Rudolf Sebastian Schittenhelm and Stephan Rinderknecht

Abstract Observers are widely used in state space control and model based fault diagnosis processes. However disturbances and model uncertainties often have large impact on the observation results regarding system states or outputs and result in reduced control or fault diagnosis performances. In rotor systems, observer design often faces two major problems: unbalances acting on the shaft are never known to full extent and in case of a rotor with large discs, gyroscopic effect results in variation of system behavior dependent on rotor rotary frequency. The predominant disturbances e.g. unbalance forces and model uncertainties caused by gyroscopic effect appear in rotor systems in a sinusoidal form with rotor rotary frequency. The influences of unbalance forces and gyroscopic effect can be considered as unknown inputs, and the signals of unknown inputs are also sinusoidal. Augmented observers that account for sinusoidal unknown inputs can be used to take advantage of this characteristic of rotor systems. The augmented observer can be applied in the control or fault diagnosis processes and can be used to estimate the distribution matrix of unknown inputs. According to the design purpose, different configurations of the augmented observer are investigated and their restrictions are discussed in this work. The performance of the augmented observer are presented and discussed with the respect to system states observation as well as fault detection and isolation.

Keywords Augmented observer · Control · Disturbance · Fault detection · Fault isolation · Gyroscopic effect · Rotor · Unbalance · Unknown input

Z. Wang (✉) · R. S. Schittenhelm · S. Rinderknecht
Institute for Mechatronic Systems in Mechanical Engineering, TU Darmstadt, Petersenstr. 30 64287 Darmstadt, Germany
e-mail: wang@ims.tu-darmstadt.de

R. S. Schittenhelm
e-mail: schittenhelm@ims.tu-darmstadt.de

S. Rinderknecht
e-mail: rinderknecht@ims.tu-darmstadt.de

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_6,

6.1 Introduction

Observers are often used in state space control and fault detection and isolation (FDI). The accuracy of the estimated system states or outputs determines the control performances or FDI performances. In rotor systems observer design often faces two major problems: disturbances acting on the shaft e.g. unbalance forces are never known to full extent and the gyroscopic effect results in rotary frequency dependent system behavior. The most widely used observers e.g. Luenberger observer and Kalman filter [8] are based on linear time invariant (LTI) systems. In cases of rotor with large discs, the gyroscopic effect cannot be neglected and the system behavior varies dependent on the rotary frequency. Thus Luenberger observer or Kalman filter with constant parameters is not applicable. If the gyroscopic matrix is modeled, an observer dependent on the rotary frequency can be constructed. But the gyroscopic matrix is often not available or is not modeled accurately. In these cases alternative methods have to be found to account for the gyroscopic effect. Actually, if a rotor model at a constant rotary frequency is used to represent the rotor system, the gyroscopic effect can be considered as model uncertainty [18]. The task is then to design an observer which is robust against disturbances and model uncertainties.

In the last decades observer design under the consideration of disturbances and model uncertainties have been widely investigated in the FDI processes [1, 2]. Wantanabe and Himelblau introduced the idea of unknown inputs observer (UIO) in [19]. Since then, different authors further developed the method and methods to describe disturbances and model uncertainties as unknown inputs are investigated. The UIO is able to estimate the system states and outputs accurately despite the presence of unknown inputs. Besides the UIO approach, other methods such as eigenstructure assignment [9, 10] and null space based methods [14, 15] are developed for FDI with the aim to eliminate the influences of unknown inputs on residuals. However in these methods the system states are not estimated accurately.

The methods introduced above are designed under the assumption that no information about the unknown inputs is available. In rotor systems excited by unbalances, the predominant disturbances e.g. unbalances are known to be sinusoidal with rotor rotary frequency. The influence of gyroscopic effect, which results in a rotary frequency dependent system behavior, also acts on the rotor system in a sinusoidal way [18]. Instead of describing the gyroscopic effect as a rotary frequency dependent term in the model, it can be represented by sinusoidal disturbance moments acting on the system [17]. Often the rotor rotary frequency can be measured and the frequency of the disturbances is known at all time. The augmented observer utilizes this characteristic of the rotor system and considers the influences of unbalances and gyroscopic effect in its design process [16].

In Sect. 6.2 a rotor test rig with unbalances and gyroscopic effect is introduced, the tests of the augmented observer are done on the basis of the finite element model of this test rig. The structure of augmented observer, which account for sinusoidal unknown inputs or disturbances, is presented in Sect. 6.3. Different configurations of the augmented observer for the purpose of states estimation and FDI can be found in this section. Isolation and identification of sinusoidal faults using augmented observers are also presented. In Sect. 6.4 a method using augmented observer to estimate unknown inputs distribution matrix, which represents how the disturbances and model uncertainties influence the system, is introduced. The feasibility tests of the observers are done and discussions of the results are given in Sect. 6.5 before a conclusion is made in Sect. 6.6.

6.2 Rotor Model

In this work the design of augmented observer is based on the model of a rotor test rig. A finite element model (FEM) is used to describe the behavior of the test rig. The feasibility tests of the augmented observer are done by means of simulation using the finite element model. Since in the reality an accurate finite element model is often not available, some limitations are considered in the observer design process.

In order to investigate rotor dynamic behavior of the low pressure shaft in a jet engine and the possibility to apply active bearing on the shaft, a test rig is available at the Institute for Mechatronic Systems in Mechanical Engineering at the Technische Universität Darmstadt. It consists of a low pressure shaft with a large disc eccentrically mounted on it, which replicates the turbine of the engine in its mechanical attributes. A passive bearing and an active double bearing are used to support the rotor. Four piezo electric actuators are installed in the active double bearing for vibration control purposes. Eight displacement sensors are installed at the rig on four planes A, B, C and D for the measurements (see Fig. 6.1).

A finite element model is built for the test rig on the basis of Timoshenko beam theory [3]. The bearings are modeled using discrete stiffness, inertia and piezoelectric elements. Due to the large disc mounted on the rotor, the gyroscopic effect is not negligible. As a result, the system dynamics are dependent on the rotary frequency Ω. The finite element model of the rotor systems is proper, thus the feed through part does not exist in the state space representation. The rotary frequency dependent model of the test rig in state space form reads:

$$\begin{aligned} \dot{x} &= A(\Omega)x + Bu + \tilde{E}\tilde{d} \\ y &= Cx, \end{aligned} \tag{6.1}$$

where x denotes the system states, u the control input, y the sensor signals and $\tilde{d}$ the disturbances, i.e. unbalance forces working on the rotor. Viscous damping is introduced to the model with a uniform damping ratio of 0.8 % for all modes. The model (6.1) is reduced for the simulations in this article to an order of 16. $A(\Omega)$, B,C,$\tilde{E}$ are system matrices with appropriate dimensions, where $A(\Omega)$ is dependent on the rotor rotary frequency Ω because of gyroscopic effect. The unbalances $\tilde{E}\tilde{d}$

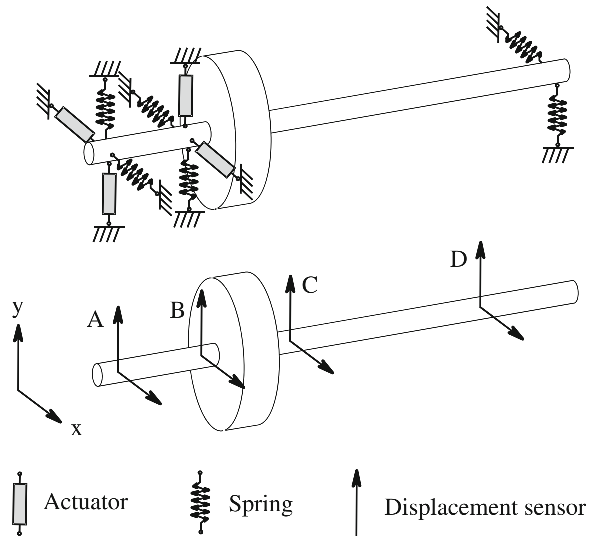


Fig. 6.1 Configuration of the rotor test rig

are simulated as randomly distributed in both axial and circumferential direction. The model (6.1) is controllable and observable for whole rotary frequency range.

In the FDI process some input faults e.g. rotor disc wear are influenced by the gyroscopic effect and periodical with the rotor rotary frequency. These faults are not simply distinguishable from the influences of gyroscopic effect and unbalances. These input faults are considered in the model:

$$\begin{aligned}\dot{x} &= A(\Omega)x + Bu + \tilde{E}\tilde{d} + Ff \\ y &= Cx,\end{aligned} \tag{6.2}$$

where f is the faults to be detected and F is its input matrix. Output faults are often simply detectable in the frequency domain and are thus not considered in this work.

The model (6.2) is used to simulate the test rig behavior in this work. In reality, usually only limited information about the system is available. Without loss of generality the model (6.2) is supposed to be unavailable for the observer design. Some limitations are considered in the design process according to the knowledge of the rotor system:

- The unbalances cannot be detected to full extent, thus the disturbance term $\tilde{E}\tilde{d}$ in Eq. (6.2) is supposed to be unknown.
- If a physical model of the rotor is available, the gyroscopic effect can be modeled as in Eq. (6.2). If the model is to be identified, the gyroscopic effect is hard to identify because of its rotary frequency dependent characteristic. Without loss of generality, it is assumed that the gyroscopic effect is not modeled and only models at specific rotary frequencies are available.

Due to the above limitations, observer design is based on the knowledge of the non-rotating rotor:

$$\begin{aligned} \dot{x} &= Ax + Bu + Ff \\ y &= Cx, \end{aligned} \tag{6.3}$$

with $A = A(0)$. The frequency dependent system matrix $A(\Omega)$ in Eq. (6.2) can be approximated by

$$A(\Omega) \approx A + \Omega A_{\Omega}. \tag{6.4}$$

The frequency dependent part ΩA_{Ω} represents the gyroscopic effect. The gyroscopic effect can be considered as moments acting on the shaft [3]. If the rotor runs at a constant frequency, gyroscopic moments and unbalance forces work on the plant in a sinusoidal form. These moments and forces are considered as unknown inputs in this article. Thus the system (6.2) simplifies to a linear time invariant system with an extra term Ed

$$\begin{aligned} \dot{x} &= Ax + Bu + Ed + Ff \\ y &= Cx, \end{aligned} \tag{6.5}$$

where $Ed = \tilde{E}\tilde{d} + \Omega A_{\Omega} x$ represents the effect of the disturbances i.e. unbalance forces and the gyroscopic moments on the system. Since d is not detected, it is referred to as unknown input and E is its distribution matrix. For the design of augmented observer only the distribution matrix E is required. Generally the matrix E can either be determined directly from the extensive knowledge of system (e.g. the gyroscopic effect and unbalance distribution in this case) or estimated by means of measurements [11–13]. Since the required information about the system for the direct determination is assumed to be unavailable, the matrix E is estimated by means of measurements in this work. Besides other estimation methods e.g. the de-convolution method, the matrix E can also be estimated using augmented observer. The configuration of the augmented observer for the estimation is introduced in Sect. 6.4.

6.3 Augmented Observer

6.3.1 Augmented State Space Model

The augmented observer introduced in this work is based on the idea of disturbance observer [6, 7], which introduces extra states in the system model to describe the influences and behavior of disturbances. If the disturbances can be described using a disturbance model:

$$\begin{aligned} \dot{x}_d &= A_d x_d \\ d &= C_d x_d, \end{aligned} \tag{6.6}$$

the augmented structure of the system reads

$$\begin{aligned} \dot{x}_B &= \begin{bmatrix} A & EC_d \\ 0 & A_d \end{bmatrix} x_B + \begin{bmatrix} B \\ 0 \end{bmatrix} u \\ y &= [C \quad 0] x_B = C_B x_B, \end{aligned} \tag{6.7}$$

with the augmented system states vector

$$x_B = \begin{bmatrix} x \\ \\ x_d \end{bmatrix} \tag{6.8}$$

The matrices A_d and C_d are respective matrices for the disturbance model and the matrix E describes how the disturbances influence the plant. If the augmented model is used for unknown inputs, the matrix E is set to unknown input distribution matrix.

An augmented observer is defined as an observer (e.g. Luenberger observer) designed on the basis of the augmented system model in Eq. (6.7). Since the disturbances are considered in the augmented system model, the estimation of the system states is theoretically accurate under the influences of the considered disturbances.

In rotor systems the unknown inputs are a set of sinusoidal signals with different amplitude and phase angles but the same frequency i.e.

$$x_B = \begin{bmatrix} \delta_1 \sin(\Omega t + \theta_1) \\ \delta_2 \sin(\Omega t + \theta_2) \\ \vdots \end{bmatrix}, \tag{6.9}$$

where Ω is the frequency of the unknown inputs and $\delta_1, \delta_2, \ldots$ and $\theta_1, \theta_2, \ldots$ are the amplitude and phase angles of the unknown inputs. Two different structures can be used for the disturbance model. If the disturbance states vector x_d in Eq. (6.6) is written as

$$x_d = \begin{bmatrix} d \\ \dot{d} \end{bmatrix}, \tag{6.10}$$

a disturbance model can be built as

$$\begin{aligned}\dot{x}_d &= \underbrace{\begin{bmatrix} 0 & I \\ -\Omega^2 I & 0 \end{bmatrix}}_{A_d} x_d \\ d &= \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{C_d} x_d.\end{aligned} \tag{6.11}$$

Another option for disturbance model uses a complementary vector of unknown inputs

$$\hat{d} = \begin{bmatrix} \delta_1 \cos(\Omega t + \theta_1) \\ \delta_2 \cos(\Omega t + \theta_2) \\ \vdots \end{bmatrix}. \tag{6.12}$$

If the disturbance states vector is set as

$$x_d = \begin{bmatrix} d \\ \hat{d} \end{bmatrix}, \tag{6.13}$$

the disturbance model is then in the form of

$$\begin{aligned}\dot{x}_d &= \underbrace{\begin{bmatrix} 0 & -\Omega I \\ \Omega I & 0 \end{bmatrix}}_{A_d} x_d \\ d &= \underbrace{\begin{bmatrix} I & 0 \end{bmatrix}}_{C_d} x_d.\end{aligned} \tag{6.14}$$

Both of the disturbance models have the same output matrix C_d. Applying the matrix C_d to Eq. (6.7), the augmented system model reads

$$\begin{aligned}\dot{x}_B &= \underbrace{\begin{bmatrix} A & \begin{bmatrix} E & 0 \end{bmatrix} \\ 0 & A_d \end{bmatrix}}_{A_B} x_B + \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{B_B} u \\ y &= [C\ 0] x_B = C_B x_B,\end{aligned} \tag{6.15}$$

The observability condition has to be hold in (6.15), so that an observer can be designed for the augmented system.

Lemma 1 If (A, C) is an observable pair and E is of full column rank, Eq. (6.15) is observable only if

$$\operatorname{rank}(E) \leq \operatorname{rank}(C), \tag{6.16}$$

i.e. rank(E) must be equal to or smaller than the number of linearly independent measurements.

Proof Assume that rank$(A) = n$ and rank$(E) = p$, according to the observability criterion of Hautus [4] the system (6.15) is observable if and only if

$$\operatorname{rank}\begin{pmatrix} \lambda_i I - A_B \\ C_B \end{pmatrix} = n + 2p \tag{6.17}$$

for all eigenvalues λ_i of $A_B \in \mathbb{R}^{(n+2p)\times(n+2p)}$. The system matrix A_B with

$$\det(\lambda_i I - A_B) = \det(\lambda_i I - A) \cdot \det(\lambda_i I - A_d) \tag{6.18}$$

has at least p pairs of eigenvalues at $\lambda_i = \pm i\Omega$ for $\det(\lambda_i I - A_d) = 0$. Thus in case of $\lambda_i = \pm i\Omega$, $\lambda_i I - A_B$ will be row rank deficient with $\operatorname{rank}(\pm i\Omega I - A_B) \leq n + p$. Thus rank$(C)$ must be greater than or equal to rank$(E) = p$, in order to satisfy condition (6.17). □

According to Lemma 1 the augmented observer can often not be applied in systems with large number of disturbances, the applicability is limited by the available sensor number. In order to apply augmented observer, there must be at least so many sensors available as the number of the considered unknown inputs; otherwise the rank of matrix E has to be reduced. One possible method to reduce the rank of matrix E is to use the technique of singular value decomposition. The matrix E can be reduced as a set of singular vectors corresponding to the most significant singular values [1, 17].

6.3.2 Augmented Observer for Control

In control theory the disturbance observer can be applied either to calculate the control inputs for disturbance compensation or for system states observation for closed loop state space control.

For the purpose of disturbance compensation an approximation of the control inputs can be estimated directly using the augmented observer. In this case the matrix E is set equal to input matrix B in the augmented state space model (6.7):

$$\begin{aligned} \dot{x}_B &= \begin{bmatrix} A & BC_d \\ 0 & A_d \end{bmatrix} x_B + \begin{bmatrix} B \\ 0 \end{bmatrix} u \\ y &= C_B x_B. \end{aligned} \tag{6.19}$$

The first row in the first equation of (6.19) can be written as

$$\begin{aligned} \dot{x} &= Ax + BC_d x_d + Bu \\ &= Ax + Bd + Bu. \end{aligned} \tag{6.20}$$

If an observer is designed on the basis of the model (6.19), the signal vector x_d can be observed and the disturbance d can be calculated directly. If the disturbances are from the system input i.e. $E = B$, the real disturbance will be observed; otherwise the model (6.19) does not match the reality and the estimation represents system input signals, which would have similar influences on the system as the disturbances. If the control inputs are set to $u = -d$, the disturbances will be compensated or attenuated.

For the purpose of system states observation for closed loop state space control, an observer can be designed on the basis of model (6.7). In this case a matrix E, that matches the reality is required. The influences of unknown inputs will then be considered in the augmented observer correctly and the estimated system states x will theoretically be accurate.

6.3.3 Augmented Observer for Fault Detection and Isolation

For fault detection purpose an augmented observer based on the augmented model (6.7) in normal Luenberger observer form can be used. The only difference to the application in the control process is that the focus of the observer design is not on the states observation but on the residual generation. Since the unknown inputs are considered in the observer, the influences of unknown inputs are included in the estimated output $\hat{y}$. The generated residual

$$r = y - \hat{y} \tag{6.21}$$

theoretically equals or approximates zero under the influences of unknown inputs i.e. the residual is then decoupled from unknown inputs with the distribution matrix E. If a fault with its input matrix $F \neq E$ takes place, the deviation on the residual is significant comparing to the influence of unknown inputs and thus easy to be detected.

In rotor systems a major class of the faults e.g. rotor disc wear are also in sinusoidal form. For fault isolation and fault identification of the sinusoidal faults, an augmented system model can be designed with the structure

$$\begin{aligned} \dot{x}_B &= \begin{bmatrix} A & [E \quad F]C_d \\ 0 & A_d \end{bmatrix} x_B + \begin{bmatrix} B \\ 0 \end{bmatrix} u \\ y &= C_B x_B. \end{aligned} \tag{6.22}$$

The respective sinusoidal fault f is then a part of the augmented states vector and can be observed directly using an augmented observer based on model (6.22). The fault isolation and identification can be done on the basis of the observed fault signals.

Note that the usage of the augmented observer for fault isolation and identification is limited by the observability condition introduced in Sect. 6.3.1. If there are not enough sensors available and different faults are assumed not to happen at the same time, a multi-observer method can be applied [5]. Using this method, multiple augmented models in the form of (6.22) can be built. In each of the augmented models only one or a few of the faults are considered, so that the observability condition holds. For each of the augmented models an observer is designed for residual generation and the multiple observers run parallel in the FDI process. If a fault takes place, the residual from the augmented observer, in which this fault is considered, will equal or approximate zero. In the meanwhile other residuals will show significant deviation from zero. In this way the fault can be isolated. The observed fault signals from the corresponding observer can be used further for fault identification.

6.4 Estimation of the Unknown Input Distribution Matrix Using Augmented Observer

If there are not enough information about the disturbances and model uncertainties available, the unknown input distribution matrix E has to be estimated by means of measurements. In rotor systems the augmented observer is applicable to estimate the distribution matrix E of the sinusoidal unknown inputs.

For the estimation of the matrix E an augmented system model is used:

$$\begin{aligned}\begin{bmatrix}\dot{x}\\ \dot{x}_{d_1}\end{bmatrix} &= \begin{bmatrix}A & [H \quad 0]\\ 0 & A_{d_1}\end{bmatrix}\begin{bmatrix}x\\ x_{d_1}\end{bmatrix} + \begin{bmatrix}B\\ 0\end{bmatrix}u\\ y &= [C \quad 0]\begin{bmatrix}x\\ x_{d_1}\end{bmatrix}.\end{aligned} \tag{6.23}$$

The disturbance model (6.11) or (6.14) can be used for A_{d_1}.

The choice of matrix H affects the observability of the augmented system (6.23). In ideal case (i.e. there are enough sensors available to fulfill the observability condition) the matrix H can be set to $H = I$. Comparing to the system model (6.15) the disturbance vector reads

$$d_1 := Ed = \begin{bmatrix}H & 0\end{bmatrix} x_{d_1}. \tag{6.24}$$

If an augmented observer is applied to (6.23), the disturbance vector d_1 can be estimated directly. The estimated vectors $d_1(k)$ for discrete time steps k are the result of a mapping of unknown inputs $d(\boldsymbol{k})$ by matrix E. The matrix E can then be estimated as a vector space in which all the vectors $d_1(k)$ lie.

In rotor systems with high system order the observability condition $\text{rank}(C) \geq \text{rank}(A)$ for $H = I$ is usually not fulfilled. Thus the matrix H has to be chosen properly, so that the Eq. (6.23) is observable and the estimation of E is as accurate as possible. Note that the frequency range of unknown inputs is limited from 0 to maximal rotor rotary frequency. In the state space representation of the rotor system the eigenforms corresponding to eigenvalues with negative imaginary parts (i.e. negative eigenfrequencies) are not excited significantly, they can be neglected without introducing too much error. The modes corresponding to eigenfrequencies that are much higher than the maximum rotary frequency usually have less influence than the low frequency ones on the system outputs. Thus influences of the high frequency modes can often also be neglected. Thus the matrix H can be chosen as

$$H = \left[e_1, e_2, \ldots, e_q\right], \tag{6.25}$$

with $q \leq \mathrm{rank}(C)$ and $e_1, e_2, \cdots, \boldsymbol{e}_q$ are eigenvectors corresponding to the eigenvalues with low positive imaginary parts.

A set of state vectors $x_{d_1}(k)$ for discrete time steps k can be observed using an observer on the basis of the augmented system model and the disturbance vectors $d_1(k)$ can be calculated as

$$d_1(k) := Ed(k) \approx \begin{bmatrix} H & 0 \end{bmatrix} x_{d_1}(k). \tag{6.26}$$

Note that $E = H$ is mathematically a solution for this problem, but for the FDI purpose it is practically not applicable. Otherwise part of the faults would be represented by the unknown inputs and can thus not be detected. In order to achieve a high fault detection rate, the matrix E is to be estimated with fewer columns. For state space control a matrix E with fewer columns means less computing time and is thus also advantageous.

For N time steps of a measurement, a set M of vectors $d_1(k)$ is calculated:

$$M = [d_1(1), d_1(2), \ldots, d_1(N)]. \tag{6.27}$$

Using singular value decomposition M is decomposed as

$$M = U\left[\,\mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n), 0\,\right] V^T. \tag{6.28}$$

The matrices U and V are left and right singular matrices and $\sigma_1, \sigma_2, \ldots, \sigma_n$ are the singular values with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$. The desired low rank approximation of E is obtained by keeping a few of the most significant singular values [1] i.e.:

$$E = U\left[\mathrm{diag}(\sigma_1, \sigma_2, \cdots, \sigma_p)\right], \tag{6.29}$$

with $p \leq n$ and $\sigma_1 \gg \sigma_{p+1} \geq \sigma_n$.

6.5 Application

6.5.1 Estimation of Unknown Input Distribution Matrix

The accuracy of the estimate of matrix E is dependent on the number of sensors applied in the system. In ideal case 16 sensors (equals the system order) are needed for an accurate estimation in the test rig. To investigate the relation between the number of sensors and estimate accuracy, different sensor configurations are tested for the estimation. As expected a 16-sensor configuration shows the best result, while a 8-sensor configuration (such as in the test rig), in which only the positive eigenfrequencies are considered does not introduce too much error in the estimation. Configurations with less than 8 sensors will reduce the accuracy significantly. In order to test the feasibility of the augmented observer for states estimation and FDI, the matrix E is estimated with high accuracy on the basis of a 16-sensor configuration.

For the purpose of system states estimation and fault detection, the gyroscopic effect is considered in the estimation and the estimate of E covers the whole rotary frequency range. The model (6.3) is used as basis model for the estimation and the system input u and fault f are not excited.

The estimation is done with measurements of stationary rotating rotor at 10 equidistant rotary frequencies. At each of the 10 rotary frequencies, a 6-column matrix E_m with $(m = 1\ldots10)$ is estimated using augmented observer introduced in Sect. 6.4 to represent both gyroscopic effect and unbalances. According to the Eqs. (6.27), (6.28) and (6.29) the estimates of E_m are proportional to the amplitude of unknown inputs. Since the rotor excitations (i.e. unbalance forces) used for the estimation are proportional to Ω^2, the 10 E_m are weighted with $1/\Omega^2$ and then combined in E_M with

$$E_M = \left[E_1^*, E_2^*, \ldots E_{10}^*\right] \tag{6.30}$$

where matrices E_m^* are weighted matrices of E_m. Using the singular value decomposition technique presented in (6.28) and (6.29), E_M is approximated by a 6-column E-matrix. This E-matrix is used for the whole rotary frequency range considered in this example.

6.5.2 Application of Augmented Observer for States Estimation

To show the advantages of augmented observer, an ordinary Luenberger observer designed for the non-rotating rotor is considered for the sake of comparison. Both ordinary Luenberger observer and augmented observer are designed on the basis of model (6.5) with 8 sensors. A constant feedback term is designed for the augmented observer, which lead to a stable observer in the whole frequency range. The poles of the both observers are set in the same region at the design point $\Omega = 0$.

Since superposition principle holds for the observed rotor system, the unbalance response without control input is taken into consideration for the sake of simplicity. The observed states are compared with real states simulated with the test rig model (6.1).

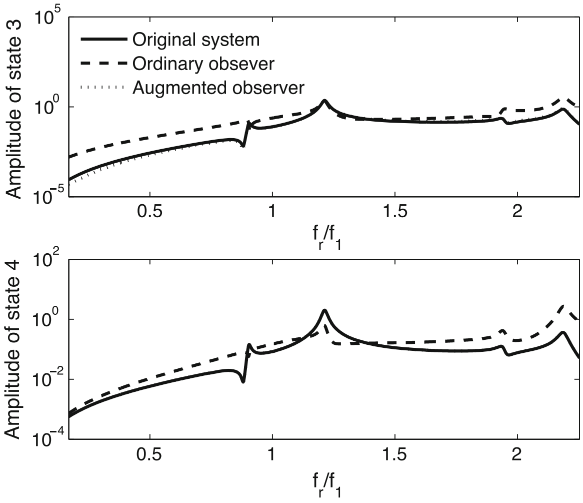


Fig. 6.2 Comparison of unbalance response on system *states* 3 and 4 in frequency domain

In Fig. 6.2 the unbalance responses of system states 3 and 4 are presented. The other states are similar and thus not presented. The excitation frequency $f_r := \Omega/2\pi$ is normalized to the first resonance frequency of the non-rotating rotor. In general the states observed by ordinary observer deviate strongly from the real states (i.e. the curve referred to original system in the figures). Thus the ordinary observer with constant parameter is not suitable for the rotor system with gyroscopic effect. The observed states by augmented observer are very accurate so that most part of the curves of the observed states and the real states coincide in Fig. 6.2. Small deviations in low frequency range can be observed in case of state 3, but the error stays in a small range and can usually be neglected for the application in state space control.

6.5.3 Application of Augmented Observer for FDI

As an example for FDI, rotor disc wear as fault is taken into consideration. The effect of this fault can be considered as an additional unbalance. Thus both unknown inputs and fault have the same frequency as rotor rotary frequency. The 8-sensor configuration of the test rig in Fig. 6.1 is used for the FDI process. For the evaluation of the results the frequency responses of both disturbances and fault on the residuals are considered. In observer based FDI, the effect of control inputs are compensated on the residual. The control inputs are thus set to 0 in the simulation.

An augmented observer based on the augmented model (6.7) is built for fault detection. The frequency responses on sensor 1 for fault free case (only excited by disturbances) and in case of fault (excited by both disturbances and fault) are presented in Fig. 6.3. Other sensors show similar results and are thus not presented. Residuals and outputs are normalized and thus have no units.

The frequency response on sensor 1 gives a brief overview about how strong the fault affects the system. The fault results mainly in a phase shift and almost no change in amplitude.

It can be seen, that in general the fault has a much stronger influence on the residual generated using augmented observer as disturbances and can thus be simply detected. Low frequency response of the fault is observed near the excitation frequency $f_r \approx 1.4 f_1$. This problem is caused by the usage of constant feedback term in observer. Since the augmented observer is dependent on rotor rotary frequency Ω and the constant feedback term is designed at a single rotary frequency, the observer has different dynamics at different rotary frequencies. Using the pole placement technique, the poles at design point are set in a region that is 10 times faster than the fastest pole of the original system. For other rotary frequencies, the poles drift away from the designed region.

As an example for fault isolation and identification, the rotor system at a constant rotary frequency is considered. It is assumed that a rotor model e.g. through identification at this frequency is known. Apart from the fault from rotor disc (fault 1), an extra unbalance in the middle of the rotor is brought to the system as fault (fault 2), Considering the rotor system at a constant rotary frequency, a E-Matrix with 2 columns is sufficient to represent the disturbances. Thus a 2-column matrix E is estimated to represent the influence of disturbances at this single rotary frequency. Based on the augmented model (6.22), an augmented observer is constructed. The faults are observed in the corresponding states x_d of the augmented states vector x_B. The time domain diagnosis is presented in Fig. 6.4.

Fault 1 takes place at 10s and fault 2 takes place at 20s. The fault amplitude is understood as amplitude of unbalance forces. The sensor signal and fault amplitudes are normalized values. A very good FDI result can be seen in Fig. 6.4.

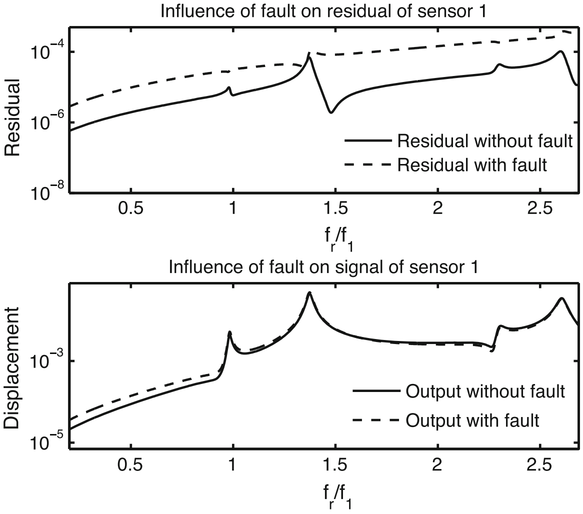
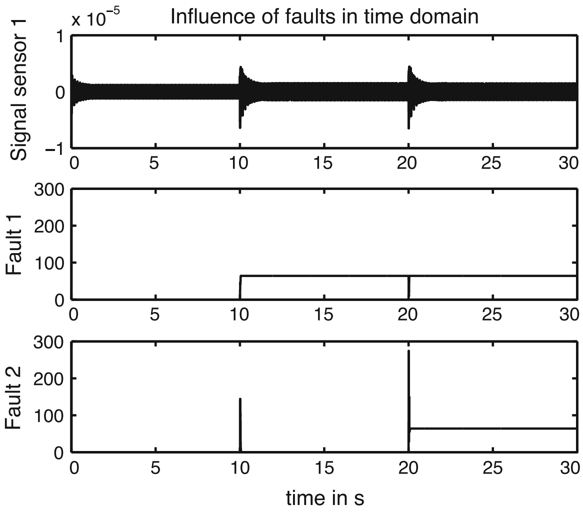


Fig. 6.3 Influence of fault on output and residual

Fig. 6.4 Observed fault amplitude by means of augmented observer



6.6 Conclusion

Augmented observer can be applied in rotor systems for the purpose of disturbance compensation, system states estimation and fault detection and isolation. Since the influences of unbalances and gyroscopic effect can be considered in the design process, the augmented observer offers better performance than ordinary observer for the application in both system states estimation and fault detection. While the unbalance forces and gyroscopic effect are considered as unknown inputs, their distribution matrix can be estimated using an augmented observer and measurements of the rotor system. The detection of unbalances and modeling of gyroscopic effect are not necessary. For faults in sinusoidal form, which often happens in rotor systems, augmented observer can also be used for fault isolation and identification. Restrictions for the application of augmented observer are discussed in this work. The feasibility of the augmented observer is proved by means of an rotor test rig model with unbalances and gyroscopic effect.

Acknowledgments This work is based on a research project in partnership with Rolls-Royce Deutschland Ltd and Co KG. and supported by German Research Foundation (DFG) within the framework of the graduate college GRK1344 “Instationäre Systemmodellierung von Flugtriebwerken”.

References

1. Chen J, Patton R (1999) Robust model-based fault diagnosis for dynamic systems. Kluwer Academic Publishers, Boston
2. Ding SX (2008) Model-based fault diagnosis techniques. Springer, Berlin

3. Genta G (2005) Dynamics of rotating systems. Springer, Berlin
4. Hautus MLJ (1969) Controllability and observability conditions of linear autonomous systems. Indagationes Mathematicae 31:443–448
5. Isermann R (2006) Fault-diagnosis systems. Springer, Berlin
6. Johnson CD (1968) Optimal control of the linear regulator with constant disturbances. IEEE Trans Autom Control 13(4):416–421
7. Johnson CD (1970) Further study of the linear regulator with disturbances - the case of vector disturbances satisfying a linear differential equation. IEEE Trans Autom Control 15(2):222–228
8. Levine W (2010) The control handbook. CRC Press Inc, Boca Raton
9. Patton R, Chen J (1991) A robust parity space approach to fault diagnosis based on optimal eigenstructure assignment. In Proceedings of the IEE Internatinonal Conference Controlõ91 (Edinburgh (1991) Peregrinus Press. 332:1056–1061
10. Patton RJ, Chen J (1991) Robust fault detection using eigenstructure assignment: A tutorial consideration and some new results. In Proceedings of the 30th IEEE Conference on Decision & Control ,Brighton, UK, pp 2242–2247
11. Patton RJ, Chen J (1993) Optimal unknown input distribution matrix selection in robust fault diagnosis. Automatica 29(4):837–841
12. Patton RJ, Chen J, Zhang HY (1992) Modelling methods for improving robustness in fault diagnosis of jet engine system. In: Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, Arizona, pp 2330–2335
13. Patton RJ, Zhang HY, Chen J (1992) Modelling of uncertainties for robust fault diagnosis. In: Proceedings of the 31st IEEE Conference on Decision and Control, Tucson, Arizona, pp 921–926
14. Varga A (2007) On designing least order residual generators for fault detection and isolation. In: Proceedings of 16th International Conference on Control Systems and Computer Science, Bucharest, Romania, pp 323–330
15. Varga A (2008) On computing nullspace bases - a fault detection perspective. In: Proceedings IFAC 2008 World Congress, Seoul, South Korea, pp 6295–6300
16. Wang Z, Schittenhelm RS, Rinderknecht S (2012) Augmented observer for fault detection and isolation (FDI) in rotor systems. In lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science, vol 1. San Francisco, USA, 24–26 Oct 2012, pp 336–341
17. Wang Z, Schittenhelm RS, Rinderknecht S (2012) Observer design for unbalance excited rotor systems with gyroscopic effect. In: Proceedings of IEEE international conference on mechatronics and automation
18. Wang Z, Wahrburg A, Rinderknecht S (2012) Consideration of gyroscopic effect in fault detection and isolation for unbalance excited rotor systems. Int J Rotating Mach 2012:14
19. Watanabe K, Himmelblau DM (1982) Instrument fault detection in systems with uncertainties. Int J Syst Sci 13(2):137–158

Chapter 7
Cooperative Tasks Using Teams of Mobile Robots

Ignacio Mas and Christopher Kitts

Abstract The coordination framework for mobile robots called cluster space control is reviewed and implemented using different robotic platforms to demonstrate specific multirobot cooperative tasks. In particular, results on multirobot object transportation and target patrolling are presented through experimental tests. Additionally, simulations on a marine oil skimming mission performed with two autonomous surface vessels are presented to illustrate the wide range of possible multirobot applications utilizing the cluster space approach. The level of abstraction introduced by this coordination framework facilitates the execution of the tasks, allowing for specification, control and monitoring of formation parameters such as position, orientation and shape of the group, instead of the positions of the individual robot members.

Keywords Cluster space control · Cooperative patrolling · Mobile robots · Multirobot applications · Object transportation · Oil skimming · Robot cooperation.

7.1 Introduction

Developments in recent years suggest that cooperative multirobot systems show great potential to improve application-specific performance by offering redundancy, increased coverage and throughput, flexible reconfiguration, or spatially

I. Mas (✉)
Consejo Nacional de Investigaciones Cientificas y Tecnicas (CONICET) and Instituto Tecnologico de Buenos Aires (ITBA), Buenos Aires, Argentina
e-mail: imas@itba.edu.ar

C. Kitts
Robotic Systems Lab, Santa Clara University, Santa Clara, CA, USA
e-mail: ckitts@scu.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_7,

diverse functionality [1]. A key aspect for mobile multirobot systems is the method by which the motions of the individual vehicles are coordinated. Centralized control approaches have been successfully demonstrated [2, 3] and have been found to be useful for material transport, regional synoptic sampling, and sensing techniques where active stimulus and/or signal reception are spatially distributed [4–6]. Such approaches, however, typically suffer from limited scalability and the need for global information. As an alternative, developments in decentralized approaches have been shown to hold great promise in addressing scalability and limited information exchange [7, 8]; such approaches often employ control strategies that are behavioral [9, 10], biologically inspired [11], optimization-based [12], market-based [13], or potential field-based [14–16].

A wide range of applications in different fields can draw benefits from formations of robots performing tasks cooperatively. Large object transportation and manipulation of objects in hazardous environments are examples of this. The group surrounds or entraps the object of interest and then transports it to a desired destination by applying pushing forces upon it [17]. Reports in the literature regarding multirobot transportation of objects show different techniques implementing this task, using approaches from behavioral-based methods [18] to lead-follower techniques [19] or potential field-based entrapment [20–22].

Another possible application in the area of security and surveillance is escorting and patrolling, where the group of robots follows a designated target of interest and protects it from hazards in the environment or keeps the target from escaping.

In the fields of marine robotics and disaster response, the autonomous containment and mechanical recovery of spilled oil from the surface of the water can eventually replace the fully human-managed operations existing nowadays. In this process, the oil is contained and concentrated with floating booms, removed from the water using skimmers, temporarily stored, and transported to shore [23]. This is a slow and manual operation and its efficiency can be greatly increased through automation.

In this article, a particular multirobot coordination approach called cluster space control framework [24] is implemented using different robotic platforms to demostrate specific tasks in a cooperative fashion. The level of abstraction introduced by the control framework facilitates the execution of the tasks, allowing for specification, control and monitoring of formation parameters such as position, orientation and shape of the group, instead of the positions of the individual robot members.

In Sect. 7.2, we review the cluster space control framework using as a case study a group of four robots, defining an appropriate set of formation parameters that represents the system, and showing how the approach is implemented in a closed-loop controller. In Sect. 7.3, the cooperative transportation of a large object is shown using a hardware experimental testbed of four non-holonomic miniature wheeled robots. The formation is piloted by a single operator using a joystick input to control in run time the position of all the robots in the group in order to push the object to the desired final position. Section 7.4 introduces a patrolling application, where a formation of three wheeled robots patrol around a moving target for protection or entrapment purposes. In Sect. 7.5, preliminary simulation results of a formation of two marine autonomous surface vessels (ASV) performing oil skimming operations is presented. The ASV are connected by a floating boom holding a U-shaped formation that collects oil from the ocean surface.

A discussion of the results reveals the advantages of using the cluster space control framework to conduct these tasks, showing in one case that only one pilot is required to perform the complex simultaneous motions of the robot formation and in the other cases that full automation or reduced task supervision or monitoring are achievable.

7.2 Multirobot Control Method

The cluster space approach to controlling formations of multiple robots was first introduced in [24]. Conceptually, the formation is represented as a fully articulated kinematic mechanism with as many degrees of freedom (DOF) as the sum of the DOF of the robot members. The first step in the implementation of the cluster space control architecture is the selection of an appropriate set of cluster space state variables. To do this, we introduce a cluster reference frame and select a set of state variables that capture key pose and geometric elements of the cluster.

The appropriate selection of cluster state variables may be a function of the application, the system's design, and subjective criteria such as operator preference. In practice, however, we have found great value in selecting state variables based on the metaphor of a virtual kinematic mechanism that can move through space while being arbitrarily scaled and articulated. This leads to the use of several general categories of cluster pose variables (and their derivatives) that specify cluster position, cluster orientation, relative robot-to-cluster orientation, and cluster shape. A general methodology for selecting the number of variables corresponding to each category given the number of robots and their DOF is described in [24]. Furthermore, an appropriate selection of cluster variables allows for centralized or distributed control architectures [25].

In this article, results with clusters of two, three, and four robots are presented. To illustrate the process, this section focuses only on the development of the method for a planar four-rover system. For such system, the robot space pose is defined as:

$$r = (x_1, y_1, \boldsymbol{\theta}_1, x_2, y_2, \boldsymbol{\theta}_2, x_3, y_3, \boldsymbol{\theta}_3, x_4, y_4, \boldsymbol{\theta}_4)^T, \tag{7.1}$$

where $(x_i, y_i, \boldsymbol{\theta}_i)^T$ defines the position and orientation of robot i. The cluster space variables can be defined as:

$$c = (x_c, y_c, \boldsymbol{\theta}_c, \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3, \boldsymbol{\phi}_4, \acute{o}, p, q, s, \beta)^T. \tag{7.2}$$

Figure 7.1 depicts the relevant reference frames for the planar four-robot problem indicating the definitions of the cluster space variables. These variables

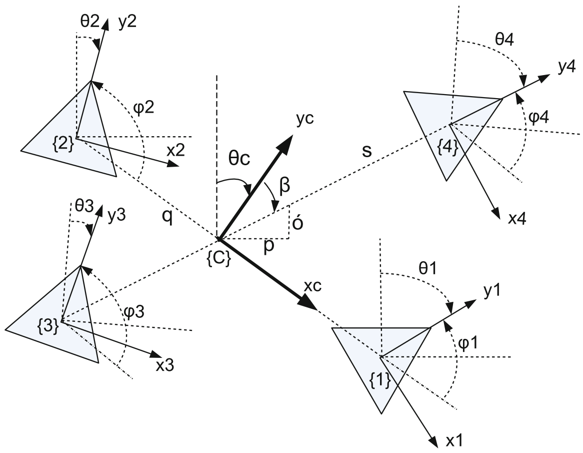


Fig. 7.1 Reference frame and cluster space variables definition for a four-robot planar system

describe the cluster position (x_c, y_c), orientation $(\boldsymbol{\theta}_c)$, and shape $(\acute{o}, p, q, s, \beta)$, as well as the relative orientation of the robots with respect to the orientation of the cluster $(\boldsymbol{\phi}_i)$. It should be noted that the resulting space definition conserves the 12 DOF–equivalent to the original system of four 3-DOF robots–therefore, the resulting system is fully articulated and any pose can be attained.

We wish to specify multirobot system motions and compute required control actions in the cluster space using the cluster state variables selected. Given that these control actions will be implemented by each individual robot (and ultimately by the actuators within each robot), we develop formal kinematic relationships relating the cluster space variables and robot space variables.

The forward position kinematics of the four-robot system are given by:

$$x_c = \frac{x_1 + x_2}{2} \tag{7.3}$$

$$y_c = \frac{y_1 + y_2}{2} \tag{7.4}$$

$$\boldsymbol{\theta}_c = atan2(x_1 - x_2, y_1 - y_2) \tag{7.5}$$

$$\boldsymbol{\phi}_i = \boldsymbol{\theta}_i - \boldsymbol{\theta}_c \qquad where\ \ i = 1, 2, 3, 4. \tag{7.6}$$

$$\acute{o} = \frac{1}{2}(y_3 + y_4) - \frac{1}{2}(y_1 + y_2) \tag{7.7}$$

$$p = \frac{1}{2}(x_3 + x_4) - \frac{1}{2}(x_1 + x_2) \tag{7.8}$$

$$q = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{7.9}$$

$$s = \sqrt{(x_3 - x_4)^2 + (y_3 - y_4)^2} \tag{7.10}$$

$$\begin{aligned} \beta = {} & atan2(x_4 - x_3, y_4 - y_3) \\ & - atan2(x_1 - x_2, y_1 - y_2), \end{aligned} \tag{7.11}$$

where $atan2(y, x)$ is the 4-quadrant arctangent [26]. The inverse position kinematics are therefore defined by:

$$x_1 = x_c + \frac{q}{2}\sin(\boldsymbol{\theta}_c) \tag{7.12}$$

$$y_1 = y_c + \frac{q}{2}\cos(\boldsymbol{\theta}_c) \tag{7.13}$$

$$\boldsymbol{\theta}_1 = \boldsymbol{\phi}_1 + \boldsymbol{\theta}_c \tag{7.14}$$

$$x_2 = x_c - \frac{q}{2}\sin(\boldsymbol{\theta}_c) \tag{7.15}$$

$$y_2 = y_c - \frac{q}{2}\cos(\boldsymbol{\theta}_c) \tag{7.16}$$

$$\boldsymbol{\theta}_2 = \phi_2 + \boldsymbol{\theta}_c \tag{7.17}$$

$$x_3 = x_c + p - \frac{s}{2}\sin(\boldsymbol{\theta}_c + \beta) \tag{7.18}$$

$$y_3 = y_c + \acute{o} - \frac{s}{2}\cos(\boldsymbol{\theta}_c + \beta) \tag{7.19}$$

$$\boldsymbol{\theta}_3 = \boldsymbol{\phi}_3 + \boldsymbol{\theta}_c \tag{7.20}$$

$$x_4 = x_c + p + \frac{s}{2}\sin(\boldsymbol{\theta}_c + \beta) \tag{7.21}$$

$$y_4 = y_c + \acute{o} + \frac{s}{2}\cos(\boldsymbol{\theta}_c + \beta) \tag{7.22}$$

$$\boldsymbol{\theta}_4 = \boldsymbol{\phi}_4 + \boldsymbol{\theta}_c, \tag{7.23}$$

We may also consider the formal relationship between the robot and cluster space velocities, $\dot{r}$ and $\dot{c}$. By differentiating the forward and inverse position kinematics, the forward and inverse velocity kinematics can easily be derived, obtaining the cluster Jacobian and Inverse Jacobian matrices, $J(r)$ and $J^{-1}(c)$ that verify $\dot{c} = J(r)\,\dot{r}$ and $\dot{r} = J^{-1}(c)\,\dot{c}$.

The particular selection of cluster space variables is not unique, and different sets of variables may be chosen following the same methodology when more convenient for a given task.

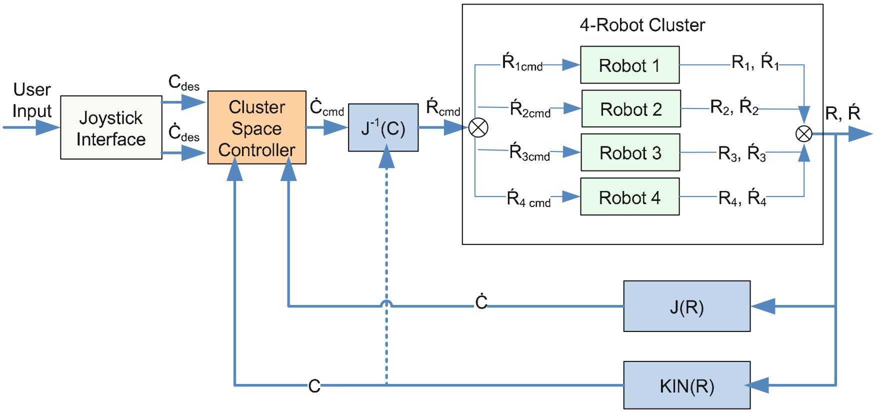


Fig. 7.2 Cluster Space Control Architecture. Desired positions and velocities are input by the pilot through a joystick interface. Control actions are computed in the cluster space and converted to the robot space through the use of the inverse Jacobian relationship. Robot sensor information is converted back to cluster space through the forward kinematics and Jacobian matrix to close the loop

Figure 7.2 presents the control architecture of the pilot-driven cluster space control of the four-robot system. The operator drives the system through a joystick interface, from which he or she can control position of the formation, orientation and shape. A cluster level PID controller compares cluster position and velocity with desired values and outputs cluster commanded velocities, which are translated into individual robot velocities through the inverse Jacobian matrix. State data from the robots are converted to cluster space information through the forward kinematics and Jacobian matrix, and fed back into the controller to close the loop.

The non-holonomic constraints given by the unicycle-like differential-drive motion of the robots effectively reduce the number of independently specified cluster pose variables to eight. As a consequence, inner-loop robot-level heading control is implemented on each robot and the cluster space controller does not regulate the four cluster parameters corresponding to yaw orientation of the robots relative to the cluster, specifically ϕ_i.

7.3 Cooperative Object Transportation

Large object transportation and manipulation of objects in hazardous environments are tasks that can benefit from the utilization of groups of mobile robots working in a cooperative fashion. The group surrounds or entraps the object of interest and then transports it to a desired destination by applying pushing forces upon it [17]. A multirobot formation control framework is necessary to coordinate the motions of the robots in the group.

Reports in the literature regarding multirobot transportation of objects show different techniques implementing this task, using approaches from behavioral-based methods [18] to lead-follower techniques [19] or potential field-based entrapments [20–22].

In this section, a testbed consisting of four non-holonomic miniature mobile robots implementing the cluster space framework is utilized to perform an object transportation task [27]. The robots used are Boe-Bots, off-the-shelf miniature rovers from Parallax Inc. The rovers are differentially-driven and carry Parallax BASIC Stamp microprocessors. These commercial units were retrofitted with X-Bee RF communication modules from Digi International Inc. Robot position and orientation were sensed using the OptiTrack Vision system from NaturalPoint Inc. The Vision-based tracking system relays position information to a central computer running a Matlab implementation of the cluster space controller. The resulting sensing accuracy is under 1 cm and the controller servo rate is 10 Hz. Desired cluster position and orientation are input by the operator with a joystick and compensation commands are sent wirelessly to the robots for actuation. Figure 7.3 shows the Boe-Bot mobile robots in their test configuration.

Figure 7.4 shows screen shots of the object transportation task performed with four robots controlled by a single operator. The robots come into contact with the square object, and the cluster follows a U-shaped trajectory specified by the operator in real-time using a joystick.

The position of the object during the experiment is shown from an overhead view in Fig. 7.5, together with the commanded cluster position that produces the object trajectory.

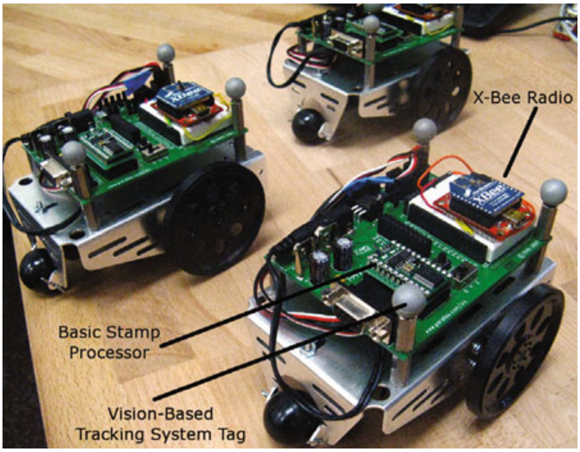


Fig. 7.3 Boe-Bot robots retrofitted with wireless communication units and vision-based tracking system tags

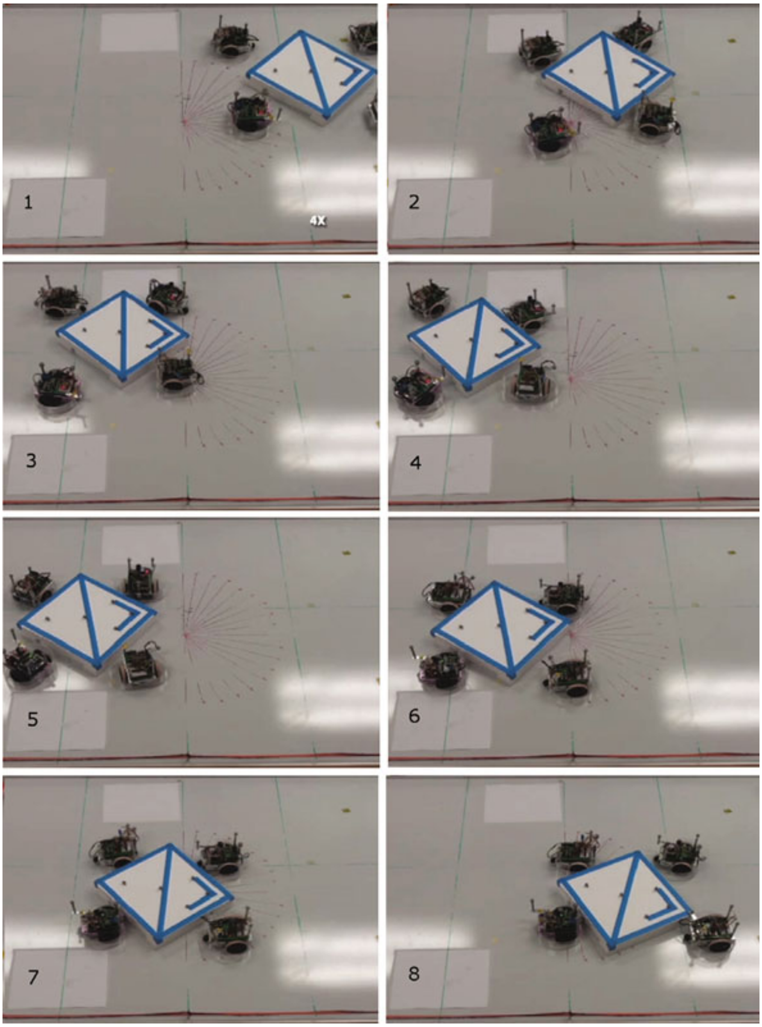


Fig. 7.4 Experimental results of the transportation of a large object. The robots entrap the object and push it following a U-shaped trajectory. The video of the experiment can be found at http://youtu.be/4eDFdNmscAg

7.4 Multirobot Escorting and Patrolling

The mission of escorting can be seen as the task of surrounding and maintaining a formation around a target whose movement is not known *a priori* but can be measured in real-time [28]. As the target moves, the formation moves to keep the target at its center, maintains the distance from the formation vehicles to the target, and evenly distributes the formation vehicles around the target. An oriented

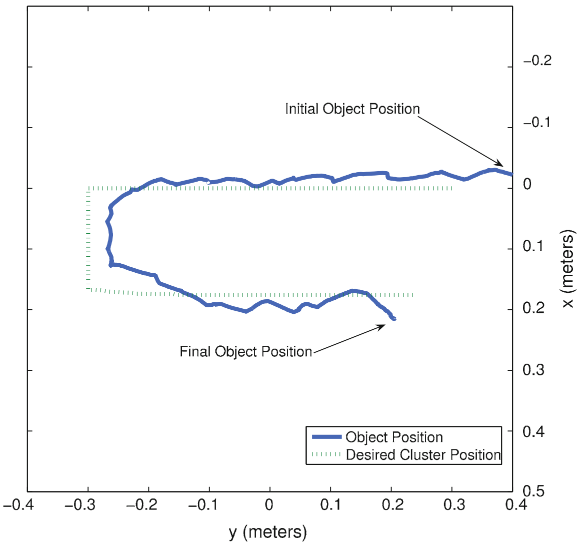


Fig. 7.5 Overhead plot of the position of the object and the commanded position of the cluster centroid for the transportation of a large object shown in Fig. 7.4. The position of the box follows the input trajectory set by the user

escorting task can be considered to be escorting when the orientation of the formation is aligned with the direction of motion of the target. The patrolling task can be defined as a further extension where the formation rotates around the target for surveillance purposes. In the literature, this is defined as a multi-agent cyclic patrolling strategy [29], where the waypoints that define the path are the initial positions of the robots relative to the target.

Entrapment may be specified similarly. In fact, from a motion control point of view, it may be implemented in precisely the same way as the escorting task, although the purpose would be to reduce target escape windows rather than preventing intrusion by external agents.

The control architecture for the escorting or patrolling mission differs from that presented in Fig. 7.2 in that the trajectory generator is fed by two new functions that are used to specify the desired cluster space trajectories. The first is a target tracking function that measures the target's position; this information is typically used to specify the centroid location for the cluster. The second is an escort/patrol profile manager that specifies the desired values for the remaining cluster variables based on the manner in which the escort/patrol function should be implemented.

Experimental tests were conducted using a cluster of three differential-drive Pioneer robots and one additional robot as a moving target. The robots are equipped with custom sensor suites based on Ultra-Wide Band (UWB) position tracking technology. The cluster definition is a triangle defined by two sides and an angle, with the cluster frame positioned at the centroid and oriented towards robot 1 as depicted in Fig. 7.6. Figure 7.7 shows the robots used in the experiment with their sensor and communication suites.

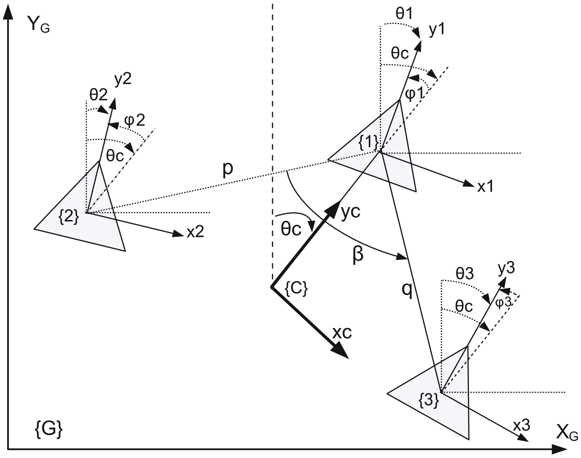
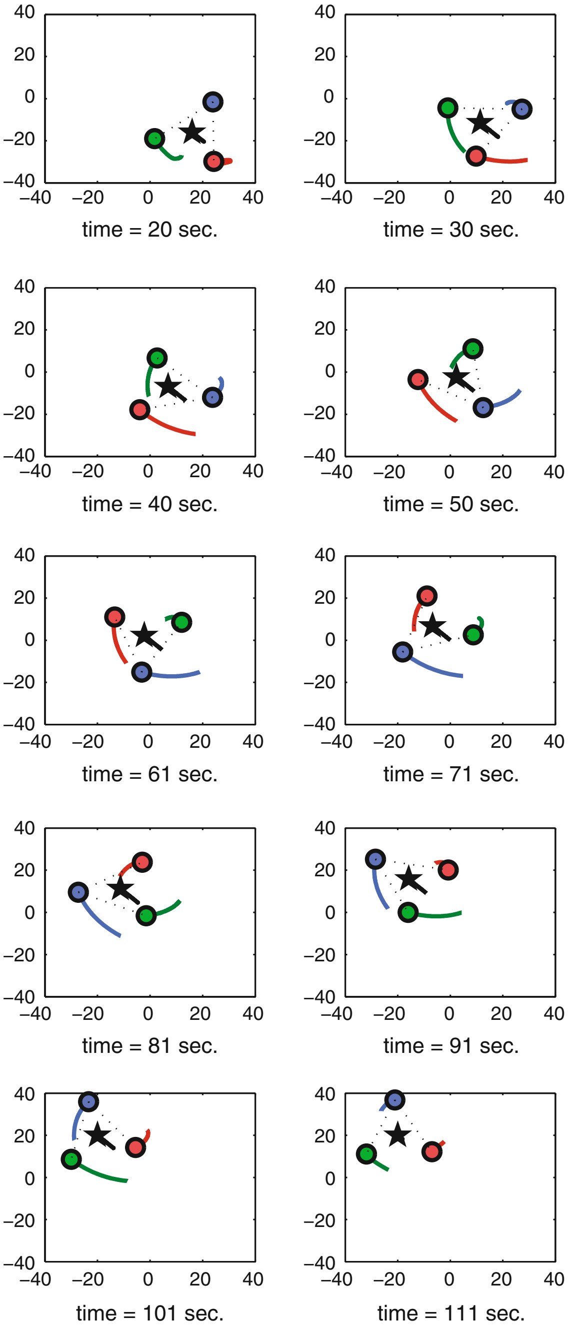


Fig. 7.6 Cluster definition for a three-robot formation where $(x_c, y_c, \boldsymbol{\theta}_c)^T$ is the cluster position and orientation, $\boldsymbol{\phi}_i$ is the yaw orientation of robot i relative to the cluster, p and q are the distances from robot 1 to robots 2 and 3, respectively, and β is the skew angle with vertex on robot 1

In the experiment, a patrolling action is performed. The target unit follows a predefined rectilinear trajectory that is unknown to the cluster. The position of the target is used as the input to the three-robot cluster x_c and y_c parameters. The formation rotates around the target keeping the relative distances. This is achieved simply by defining a linear trajectory for the cluster orientation parameter $\boldsymbol{\theta}_c$ [30]. Figure 7.8 shows the motion of the individual robots resulting from this specification. The robots get to their position and start tracking the target while patrolling around it. The tracking mean square errors (difference between target and cluster centroid positions) in x and y, after the initial transients are over, are 0.042 and 0.027 m, respectively.

Fig. 7.7 Pioneer robots from Adept Mobilerobots LLC., with GPS receivers, digital compass and wireless communication units

Fig. 7.8 Patrolling experiment using a formation of three wheeled robots. The target (*star*) moves from the *bottom right* to the *top left* and stops. The formation follows the target while rotating around it. Axes units represent meters



7.5 Autonomous Oil Skimming

Ocean oil skimming operations currently have a very limited impact in disaster recovery. The reason being that skimming (whether with a floating boom deployed from one ship, or via the manipulation of a flexible skimmer using two or more ships) is a slow, human-intensive operation. Given the widespread concern at the use of certain chemicals for dispersion, increased efficiency in skimming operations could lead to positive change in the way cleanup is performed in the future [31]. This task is similar to that of marine caging, where a floating device is captured and shepherded to a designated position [32].

We propose the use of the cluster space control architecture in order to control a formation of two vessels performing oil cleanup operations. Cluster space control provides a simple method of specifying and controlling groups of robots and sets a direction towards fully autonomous oil skimming operations.

Proper selection of the cluster space variables arranges two autonomous surface vessels connected by a floating boom in a U-shaped formation desired for oil skimming, as decipted in Fig. 7.9 This particular specification has the additional benefit of allowing direct control of the inter-robot distance spanned by the boom. The cluster space definition for the two ASV system is shown in Fig. 7.10.

Computer simulations in Matlab/Simulink include the modeling of two ASV and a floating boom connected to them. The dynamic models of the boats are based on experimental data [33], and the boom is modeled as discrete segments derived from first principles [31] and implemented in Matlab SimMechanics.

Figure 7.11 shows the dynamic response of the floating boom as the vessels start traveling. The model has the capability to include disturbances such as currents and wind.

A simulation run of a formation of two ASV performing a lawn moving pattern is shown in Fig. 7.12. This behavior reflects the required motions for an autonomous oil skimming operation.

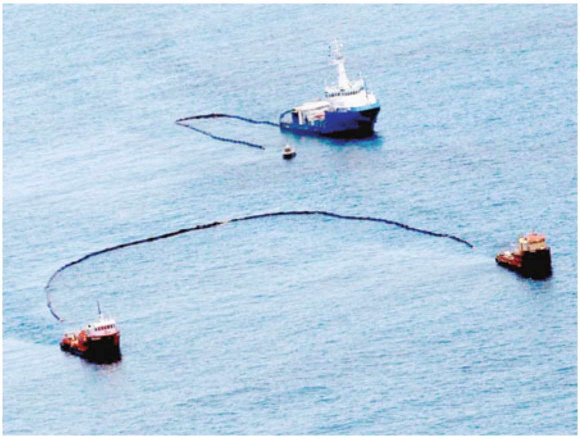


Fig. 7.9 Skimmer boats work to collect oil near the BP Deepwater Horizon oil spill on June 25, 2010 in the Gulf of Mexico off the coast of Louisiana. (www.abcactionnews.com)

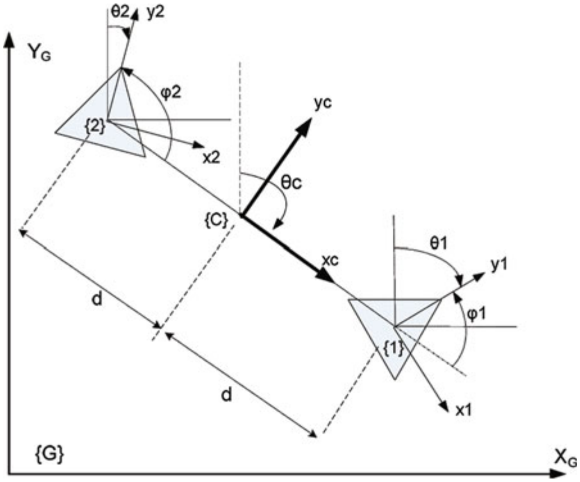


Fig. 7.10 Cluster space variables definition for a two robot system

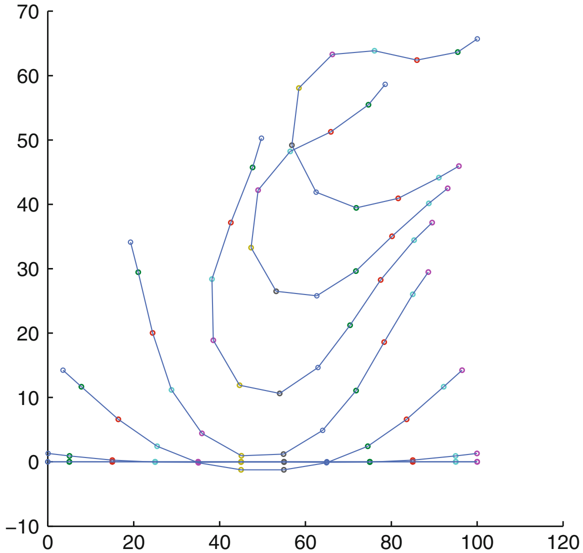
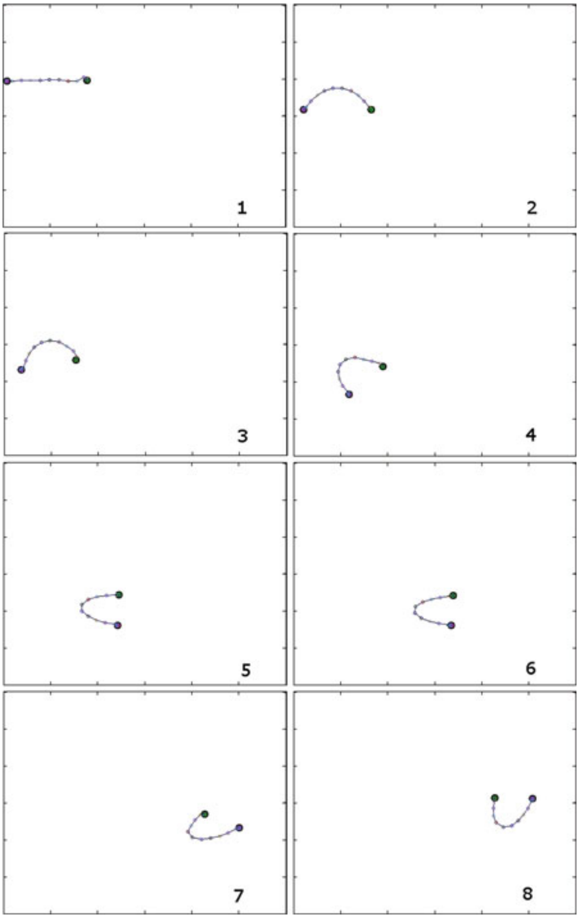


Fig. 7.11 Oil skimming boom dynamic model. Axes units represent meters

7.6 Discussion

Applying cluster space control to object transportation and manipulation tasks has some advantages and shortcomings compared to other methods found in the literature. For example, our proposed method allows for the robots to simultaneously move in the workspace, with no need to take turns to perform discrete actions as in [18]. In contrast to lead-follower methods [19, 34], our approach allows for a higher level of abstraction in the definition of the group. Consequently, specifying

Fig. 7.12 Formation of two robots performing an oil skimming task. Tick mark intervals represent 50 m



motions like 'rotate formation' or 'increase formation size' can be achieved through a smaller number of control variables. As a drawback, a higher level of communication is needed and robustness to robot failure decreases. Compared to potential field-based approaches [20], cluster space control does not need *a priori* knowledge of the shape of the object to be transported, a requirement to create a suitable potential function. The operator can modify the shape and orientation of the formation in run time to perform the task.

The proposed coordination approach allows for different sets of cluster space variables to be defined for a given formation. The operator can therefore benefit from a particular selection of variables that responds to the control needs of the task. In the patrolling example presented in this article, the selection of the centroid of the robots as the cluster position simplifies the task of surrounding the object and keeping each robot at a constant distance from it, as the position of the cluster centroid and that of the tracked target coincide, and the rotation around the target is performed by simply changing the orientation of the formation around the centroid. Variations on the cluster configuration allow for alternative protection tasks such as marine asset guarding [33].

Automated oil skimming is a task that holds great promise for the future but there are several challenges to be faced. Adjusting and verifying the dynamic model of the floating boom and its interaction with the ASV need to be addressed, as well as implementing controllers able to compensate for such disturbances. Experimental testing with scaled down vessels connected by a floating rope are currently being developed [35].

7.7 Conclusion

The cluster space framework was reviewed and illustrated with a four-robot example case, showing the cluster variables selection process and the kinematic transforms. Then, three examples of possible field applications using cooperative formations of mobile robots were proposed. First, the cluster space control framework was applied to a four-robot formation for transportation of a large object. Given the level of control abstraction introduced by the cluster space variables, the experiment show that a single pilot or operator can effectively command and monitor the position of the robots in the group in order to cooperatively accomplish the task. Then, experimental data of a patrolling task was shown using three wheeled robots, where the framework allows for simple mission specification resulting in complex robot behaviors. Lastly, simulations of two ASV carrying a floating boom and performing oil skimming operations were presented to illustrate the range of possible applications and trailblaze new directions on fielding multirobot systems that can solve concrete engineering problems.

Acknowledgments The authors gratefully thank Steve Li and Thomas Adamek for their help developing and maintaining the experimental testbeds and Mike Rasay for improving the Boe-Bot mechanical design. This work has been sponsored through a variety of funding sources to include Santa Clara University Technology Steering Committee grant TSC131 and National Science Foundation Grant No. CNS-0619940. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. Kitts C, Egerstedt M (2008) Design, control, and applications of real-world multirobot systems [from the guest editors]. Rob Auto Mag IEEE 15(1):8–8. doi:10.1109/M-RA.2007.914989
2. Yamaguchi H, Arai T (1994) Distributed and autonomous control method for generating shape of multiple mobile robot group. In: Intelligent robots and systems, IROS '94.

Proceedings of the IEEE/RSJ/GI international conference on, vol 2:800–807. doi:10.1109/IROS.1994.407547

3. Tan KH, Lewis M (1996) Virtual structures for high-precision cooperative mobile robotic control. In: Intelligent robots and systems. IROS, Proceedings of the IEEE/RSJ international conference on, vol 1:132–139. doi:10.1109/IROS.1996.570643
4. Hashimoto M, Oba F, Eguchi T (1993) Dynamic control approach for motion coordination of multiple wheeled mobile robots transporting a single object. Intelligent robots and systems '93, IROS '93. Proceedings of the 1993 IEEE/RSJ international conference on, vol 3, pp 1944–1951. doi:10.1109/IROS.1993.583900
5. Rus D, Donald B, Jennings J (1995) Moving furniture with teams of autonomous robots. In: Intelligent Robots and Systems, Proceedings. 1995 IEEE/RSJ International Conference on, vol 1, pp 235–242. doi:10.1109/IROS.1995.525802
6. Tang CP, Bhatt R, Abou-Samah M, Krovi V (2006) Screw-theoretic analysis framework for cooperative payload transport by mobile manipulator collectives. Mechatron IEEE/ASME Trans 11(2):169–178. doi:10.1109/TMECH.2006.871092
7. Siljak D (1991) Decentralized control of complex systems. Academic, New York
8. Tychonievich L, Cohoon J (2012) Coalescing swarms of limited capacity agents: Meeting and staying together (without trust). IAENG Int J Comput Sci 39(3):254–260
9. Balch T, Hybinette M, (2000) Behavior-based coordination of large-scale robot formations. MultiAgent Systems, (2000) Proceedings. Fourth international conference on, pp 363–364. doi:10.1109/ICMAS.2000.858476
10. Flinn E (2005) Testing for the 'boids'. Aerosp America 43(6):28–29
11. Murray RM (2007) Recent research in cooperative control of multi-vehicle systems. J Dyn Syst Meas Control 129(5):571–583
12. Dunbar W, Murray RM (2006) Distributed receding horizon control for multi-vehicle formation stabilization. Automatica 42(4):549–558
13. Zhu W, Choi S (2011) A closed-loop bid adjustment approach to dynamic task allocation of robots. Eng Lett 19(4):279–288
14. Leonard N, Fiorelli E (2001) Virtual leaders, artificial potentials and coordinated control of groups. Decision and Control. Proceedings of the 40th IEEE Conference on, vol 3, pp 2968–2973. doi:10.1109/.2001.980728
15. Ogren P, Fiorelli E, Leonard N (2004) Cooperative control of mobile sensor networks:adaptive gradient climbing in a distributed environment. Auto Control IEEE Trans 49(8):1292–1302. doi:10.1109/TAC.2004.832203
16. Justh EW, Krishnaprasad PS (2004) Equilibria and steering laws for planar formations. Sys Control Lett 52:25–38
17. Pereira GAS, Kumar V, Spletzer J, Taylor CJ, Campos MFM (2002) Cooperative transport of planar objects by multiple mobile robots using object closure. Exp Rob VIII, 275–284
18. Mataric M, Nilsson M, Simsarian K (1995) Cooperative multi-robot box-pushing. Intelligent robots and systems. IEEE/RSJ international conference on, In, pp 556–561
19. Wang Z, Takano Y, Hirata Y, Kosuge K (2007) Decentralized cooperative object transportation by multiple mobile robots with a pushing leader. Distrib Auton Rob Sys 6:453–462. doi:10.1007-978-4-431-35873-2-44
20. Song P, Kumar V (2002) A potential field based approach to multi-robot manipulation. In: Robotics and automation. IEEE international conference on, vol 2:1217–1222. doi:10.1109/ROBOT.2002.1014709
21. Fink J, Hsieh M, Kumar V (2008) Multi-robot manipulation via caging in environments with obstacles. In: Robotics and automation, 2008. ICRA 2008. IEEE international conference on, pp 1471–1476 (2008). doi:10.1109/ROBOT.2008.4543409
22. Yamashita A, Ota J, Arai T, Ichikawa K, Kamata K, Asama H (2001) Cooperative manipulation and transportation of a large object by multiple mobile robots. In: Asama H, Inoue H (eds) Intelligent autonomous vehicles 2001, pp 375–380
23. Tebeau P (2003) Us coast guard oil spill response research & development program, a decade of achievement. Tech. rep, DTIC Document

24. Kitts CA, Mas I (2009) Cluster space specification and control of mobile multirobot systems. Mechatron IEEE/ASME Trans 14(2):207–218. doi:10.1109/TMECH.2009.2013943
25. Mas I, Kitts C (2010) Centralized and decentralized multi-robot control methods using the cluster space control framework. Advanced intelligent mechatronics (AIM), 2010 IEEE/ ASME international conference on, pp 115–122. doi:10.1109/AIM.2010.5695768
26. Craig J (2005) Introduction to robotics. Mechanics and control, 3rd edn. Pearson Prentice Hall, NJ
27. Mas I, Kitts C (2012) Object manipulation using cooperative mobile multi-robot systems. Lecture Notes in Engineering and Computer Science: Proceedings of the world congress on engineering and computer science 2012. WCECS 2012:324–329
28. Antonelli G, Arrichiello F, Chiaverini S (2007) The entrapment/ escorting mission for a multi-robot system: theory and experiments. Advanced intelligent mechatronics, 2007 ieee/ asme international conference on, pp 1–6. doi:10.1109/AIM.2007.4412504
29. Chevaleyre Y (2004) Theoretical analysis of the multi-agent patrolling problem. Intelligent agent technology, IEEE / WIC / ACM international conference on, pp 302–308 http://doi.ieeecomputersociety.org/10.1109/IAT.2004.1342959
30. Mas I, Li S, Acain J, Kitts C (2009) Entrapment/escorting and patrolling missions in multi-robot cluster space control. Intelligent robots and systems. IEEE/RSJ international conference on, In, pp 5855–5861
31. Bhattacharya S, Heidarsson H, Sukhatme G, Kumar V (2011) Cooperative control of autonomous surface vehicles for oil skimming and cleanup. In: Robotics and automation (ICRA), 2011 IEEE international conference on, pp 2374–2379. IEEE (2011)
32. Arrichiello F, Heidarsson H, Chiaverini S, Sukhatme G (2010) Cooperative caging using autonomous aquatic surface vehicles. In: Robotics and automation (ICRA), 2010 IEEE international conference on, pp 4763–4769. IEEE (2010)
33. Mahacek P, Kitts C, Mas I (2012) Dynamic guarding of marine assets through cluster control of automated surface vessel fleets. Mechatron IEEE/ASME Trans 17(1):65–75. doi:10.1109/TMECH.2011.2174376
34. Spletzer J, Das A, Fierro R, Taylor C, Kumar V, Ostrowski J (2001) Cooperative localization and control for multi-robot manipulation. In: Intelligent robots and systems, 2001. Proceedings. 2001 IEEE/RSJ international conference on, vol 2, pp 631–636 vol. 2. doi:10.1109/IROS.2001.976240
35. Neumann M, Adamek T, Mas I, Kitts C (2012) Extension of cluster space control for 3-vessel oil skimming. In: Proceedings of the 2012 ASME/JSME joint international conference on micromechatronics for information and precision equipment (MIPE2012) (2012)

Chapter 8
An Optimized, Authenticated Key Distribution Protocol for Optical Channels

Sara Abozied, Hassan Elkamchouchi, Yasmine Abouelseoud and Refaat El-Attar

Abstract In this paper, a key distribution protocol based on the integration of both classical and quantum cryptography is developed. Quantum physics laws ensure detection of eavesdropping over optical channels, while classical cryptography provides convenient techniques that enable efficient user authentication and prevent denial of previous commitments. The proposed scheme is based on the RSA-TBOS signcryption scheme to achieve the combined functionality of a digital signature and encryption in an efficient manner. It therefore offers four security services: privacy, authenticity, data integrity and non-repudiation. The ciphertext is converted to binary bits and then to qubits. The protocol involves a small number of rounds and the number of keys stored at the user side is kept minimal promoting its use in large networks. A comparative study with other schemes in literature revealed that it represents a remarkably good trade-off between the storage as well as bandwidth requirements and the security services offered.

Keywords Authentication · Classical cryptography · Digital signature · Quantum cryptography (QC) · RSA-TBOS signcryption · Session key

S. Abozied (✉)
Engineering Mathematics Department, Alexandria University, Alexandria, Egypt
e-mail: saraabozied@gmail.com

H. Elkamchouchi
Electrical Engineering Department, Alexandria University, Alexandria, Egypt
e-mail: helkamchouchi@gmail.com

Y. Abouelseoud
Engineering Mathematics Department, Alexandria University, Alexandria, Egypt
e-mail: yasmine.abouelseoud@gmail.com

R. El-Attar
Engineering Mathematics Department, Alexandria University, Alexandria, Egypt

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_8,

8.1 Introduction

Security has become a big concern in wired and wireless networks. The characteristics of networks pose both challenges and opportunities in achieving security goals, such as confidentiality, authentication, integrity, availability, access control, and non-repudiation. Cryptographic techniques are widely used for secure communications. All security services are somewhat based on sharing a secret key among the communicating parties. Symmetric (private) key cryptosystems are usually computationally more efficient than asymmetric (public) key cryptosystems. Thus, symmetric key encryption schemes are used to encipher the bulk text, while asymmetric key mechanisms are used for enciphering the secrets to be shared among communicants. Asymmetric cryptosystems provide efficient mechanisms for authentication and non-repudiation.

Classical digital bits are easy to copy without leaving any traces behind. This motivated the use of microscopic objects, such as photons, for encoding the transmitted data. This is because quantum physics laws, such as Heisenberg's uncertainty principle and the no-cloning theorem, guarantee that any observation of such microscopic objects will inevitably change its state and hence eavesdropping can be detected [1, 2].

An interesting solution to the delicate problem of distribution of keys met in cryptography is the use of the laws of quantum physics. Quantum Cryptography (QC) protocols are used to carry out the task of exchanging keys with great security. Quantum cryptography has been proven secure even against the most general attack allowed by the laws of physics and is a promising technology for adoption in realistic cryptographic applications [1]. However, eavesdropping is detected after it takes place and this explains the use of quantum cryptography mainly for exchanging insensitive information, such as keys which can be reset without imposing any security hazards.

The bit is the fundamental unit of classical computation and classical information. Quantum computation is built upon an analogous concept, the quantum bit, or qubit for short. Just as a classical bit has a state—either 0 or 1—a qubit also has a state. Two possible states for a qubit are the states $|0>$ and $|1>$, which correspond to the states 0 and 1 for a classical bit. The difference between bits and qubits is that a qubit can be in a state other than $|0>$ or $|1>$. It is possible to form linear combinations of states, often called superpositions:

$$|\psi> = \alpha|0> + \beta|1>$$

where the numbers α and β are complex numbers satisfying

$$|\alpha|^2 + |\beta|^2 = 1.$$

A more important task to be done prior to communication is authentication that guarantees that the origin of the message is genuine because, if a malicious user masquerades as a legitimate user, the key distribution schemes and encryption schemes will be easily compromised. In situations where there is not complete trust between the sender and the receiver, something more than authentication is needed which is the digital signature. The digital signature is analogous to the handwritten signature. It must have the following properties:

- It must verify the author and the date and time of the signature.
- It must authenticate the contents at the time of the signature.
- It must be verifiable by third parties, to resolve disputes.

Thus, the digital signature function includes the authentication function. The straightforward use of public-key encryption provides confidentiality but not authentication. The source uses the public key PU_b of the destination to encrypt M. Because only B has the corresponding private key PR_b, only B can decrypt the message. This scheme provides no authentication because any opponent could also use B's public key to encrypt a message, claiming to be A.

To provide authentication, A uses its private key to encrypt the message, and B uses A's public key to decrypt it. This provides authentication using the same type of reasoning as in the symmetric encryption case: The message must have come from A because A is the only party that possesses PR_a and therefore the only party with the information necessary to construct the ciphertext that can be decrypted with PU_a.

To provide both confidentiality and authentication, A can encrypt M first using its private key, which provides the digital signature, and then using B's public key, which provides confidentiality. The disadvantage of this approach is that the public-key algorithm, which is complex, must be exercised four times rather than two in each communication [3].

In this paper, a key distribution protocol based on transmission of qubits is proposed to achieve: privacy, authenticity, data-integrity and non-repudiation. The protocol relies on the use of the classical signcryption cryptographic primitive to achieve these goals in an efficient way with regard to bandwidth and storage requirements. Moreover, the detection of eavesdroppers is achieved according to laws of physics through encoding the transmitted data on microscopic objects such as photons.

Our primary goal is to construct a scheme such that the number of keys stored per user and the number of rounds are kept minimal. Additionally, no third party knows the shared session key in the proposed scheme.

8.2 Related Work

In this section, several quantum-based key distribution and authentication schemes in literature are reviewed.

8.2.1 BB84 Protocol

The BB84 protocol was first introduced in 1984 by Charles Bennett of IBM Research and Gilles Brassard of the University of Montreal [4]. It suffers from several problems. It is susceptible to man-in-the-middle attack, about 50 % of the bandwidth is wasted, its number of rounds is high and there is no authentication. Moreover, it doesn't withstand the beam splitting attack.

In spite of this, it is still widely used and has become standard. It is based on Heisenberg's uncertainty principle. The BB84 protocol uses polarized photons. Alice sends polarized photons, referenced to one of two different orthogonal base sets (i.e., {horizontal, vertical} or {+ 45°, −45°}), and Bob observes the received photon, randomly choosing one of the two bases. After a certain amount of data is transmitted, Alice and Bob determine which data bits should be discarded by exchanging information about the bases they used for polarizations and measurements using a classical channel. They keep the rest of the data bits after sifting as the key for future use. Hence, the length of the key is in the order of half the number of the bits transmitted [5].

8.2.2 An Authentication Protocol Using Quantum Entangled States

Entanglements are one of the most mysterious aspects of quantum mechanics and are the source of the power of quantum computation. Suppose there are two qubits, the first in the state $\alpha_0|0> + \alpha_1|1>$ and the second in the state $\beta_0|0> + \beta_1|1>$. The joint state of the two qubits is the tensor product of the two; that is, $\alpha_0\beta_0|00> + \alpha_0\beta_1|01> + \alpha_1\beta_0|10> + \alpha_1\beta_0|10> + \alpha_1\beta_1|11>$. Given an arbitary system of two qubits, in general, one cannot specify the state of each individual qubit. The two qubits are most probably entangled and cannot be decomposed into the states of the individual qubits. For example, consider the Bell state

$$|\psi> = \frac{|00> + |11>}{\sqrt{2}}$$

Entangled quantum states have been used in several quantum authentication protocols such as the work in [6]. Each of the communicating parties is in possession of one of the two particles that constitute together the two qubits system. Usually, n entangled states are shared to reduce the probability of fraudulent impersonation attacks.

8.2.3 A Quantum Authentication Protocol Using Superposition States

It is a two-party authentication protocol. To hide transmitted data from unauthorized users, this protocol uses quantum superposition states instead of quantum entangled states. Authenticating a specific user within a group of many users using quantum entangled states is a difficult task. This protocol works well under the assumption that both parties already share a secret key (k_a). Furthermore, it was shown that the superposition states can be realized using current technologies (e.g., linear polarizers and Faraday rotators). This protocol is secure against the beam splitting attack and the intercept/resend attack. But, this protocol in the multi-user setting will involve the storage of a large number of pre-shared keys per user in the network [5].

8.2.4 A Three-Party Quantum Authentication Protocol Using Superposition States

The objective of this protocol is to let participants share a different session key for each new session while providing authentication, both implicitly and explicitly. To hide transmitted data from unauthorized users, this protocol uses quantum superposition states instead of entangled states as the previous protocol. This protocol consists of three phases. In the first phase, the participants are implicitly authenticated using the trusted center (TC). In the second phase, a session key is established between the two participants. Even the trusted center cannot listen to the secure communication between the participants because the session key shared between the participants is hidden from the trusted center. In the third phase, the participants of the communication are mutually authenticated to each other in an explicit way, as detailed in [7]. This protocol is secure against the beam splitting attack and the intercept/resend attack. The presence of a TC resolves the problem of storing a huge number of pre-shared keys in the multi-user case in a large network. The problem of this protocol is that the number of its rounds is high.

8.2.5 AMNI'09 Protocol

In this protocol, a session key is transmitted to the users by a trusted center, which generates the public key and the private key for each user in the registration phase. The trusted center uses the Rivest, Shamir, Adleman (RSA) asymmetric algorithm to encrypt the session key which is then converted to qubits for transmission to each user who wishes to participate in a communication session. In this protocol, security is achieved but the authentication is weak; i.e. not achieving non-repudiation [2]. In this protocol, a trusted center is used and it knows the session key so it will know the transmitted message.

8.3 Signcryption

Signcryption is a combination of a digital signature algorithm and an encryption algorithm. We review a signcryption scheme based on the RSA trapdoor one-way function [8]. An attractive feature of this scheme is that it offers non-repudiation and data integrity check in a very simple manner. The size of the result of this signcryption scheme is about half the size of a signed and encrypted message using standard RSA techniques. For this reason, they gave it the name “Two Birds One Stone (TBOS)”; that is, signcryption at the cost of encryption.

8.3.1 Key Parameters

- k: Even positive integer.
- Sender (Alice’s) RSA public and private key pair: (N_A, e_A) and (P_A, Q_A, d_A), respectively, where P_A and P_B are the two prime factors of N_A
- Receiver (Bob’s) RSA public and private key pair: (N_B, e_B) and (P_B, Q_B, d_B), respectively, where P_E and Q_E are the two prime factors of N_B.

Note: We must have $|N_A| = |N_B| = k$.

- Two hash functions H and G, where $H{:}\{0,1\}^{n+k_0} \rightarrow \{0,1\}^{k_1}$ and $G{:}\{0,1\}^{k_1} \rightarrow \{0,1\}^{n+k_0}$ and $k = n + k_0 + k_1$, with 2^{-k_0} and 2^{-k_1} being negligible.

8.3.2 TBOS Signcryption Module

When Alice signcrypts a message $M \in \{0,1\}^n$ for Bob, she performs:

1. $r \leftarrow \{0,1\}^{(k_0)}$
2. $w \leftarrow H(M \,||\, r)$
3. $s \leftarrow G(w) \oplus (M||r)$
4. If $s \,||\, w \; > N_A$ goto 1
5. $c' \leftarrow (s \,||\, w)^{d_A} \bmod N_A$
6. If $c' \; > N_B, c' \leftarrow c' - 2^{k-1}$
7. $c \leftarrow c'^{e_B} \bmod \; N_B$
8. Send c to Bob

8.3.3 TBOS Unsigncryption Module

When Bob unsigncrypts a cryptogram received from Alice, he performs:

1. $c' \leftarrow c^{d_B} \bmod N_B$
2. If $c' > N_A$, reject
3. $\mu \leftarrow c'^{e_A} \bmod N_A$
4. Parse μ as $(s \parallel w)$
5. $M \parallel r \rightarrow G(w) \oplus s$
6. If $H(M \parallel r) = w$, return M
7. $c' \leftarrow c' + 2^{k-1}$
8. If $c' > N_A$, reject
9. $\mu \leftarrow c'^{e_A} \bmod N_A$
10. Parse μ as $(s \parallel w)$
11. $M \parallel r \leftarrow G(w) \oplus s$
12. If $H(M \parallel r) \neq w$, reject
13. Return M.

It can be shown that given a valid signcrypted text, the unsigncryption algorithm returns the original plaintext.

8.4 The Proposed Scheme

In our protocol, integration of quantum cryptography for secure optical transmission and classical cryptography for identity authentication is considered. In many of the existing quantum key distribution schemes, the number of communication rounds is large and the identity of the user is not verified. The proposed protocol is an enhancement of the work in [2], where the integrity of the key transmitted is verified. A preliminary version of this work appeared in [9].

The ultimate goal of this protocol is that transmitter and the receiver share an authenticated session key ‘SK’, which is an n-bit random number.

In what follows, the steps of the proposed protocol are provided. We assume that every participant shares a secret key with the trusted center in advance. Let $K_{A,T}$ be the key shared between Alice and the TC, and $K_{B,T}$ be the key shared between Bob and the TC. Those keys serve for the mutual authentication between the trusted center and each of the communicating parties.

Let h(K,M) be a hash value of a message M with key K, generated using a cryptographic hash function (e.g., SHA-1 or MD5).

Step 1: Registration phase

The private key and the public key are generated using the RSA algorithm for each user.

(1) Choose two large prime numbers P and Q.
(2) Compute N = P*Q.
(3) Choose e (less than N) such that e and A = (P–1)(Q–1) are relatively prime (having no common factor other than 1), the public key is (N, e).
(4) Choose d such that (e*d) mod [(P–1)(Q–1)] is equal to 1, the private key is (A,d).

The public key of each user can be openly exchanged and the user's private key is kept secret. Each user obtains a certificate from the TC (*cert*) for its public key providing the link between the user's identity and its key.

Step 2: Sharing a random number for bases synchronization

(1) The TC generates a random number r. The transmitter and the receiver synchronize their quantum polarization bases in step 5 according to this pre-shared random number. Then, the TC computes:

$$X = h(K_{A,T}, r) \oplus (U_A || U_B)$$

$$Y = h(K_{B,T}, r) \oplus (U_B || U_A)$$

where || indicates the concatenation of the bit strings and U_X indicates the identifier of the participant X including a public key and its associated certificate.
(2) Now, r || X is encrypted with the pre-shared key $K_{A,T}$ using a suitable quantum block encryption scheme as that in [10] and the result is transmitted to Alice (the transmitter) over a quantum channel. Similarly, r || Y is encrypted with the pre-shared key $K_{B,T}$ and the result is transmitted to Bob (the receiver) over another quantum channel.
(3) The transmitter decrypts and measures the received qubits. It computes a hash value using $K_{A,T}$ and r, and obtains the values of $U_A || U_B$. Then, it verifies the values of U_A and U_B.
(4) The receiver decrypts and measures the received qubits. It computes a hash value using $K_{B,T}$ and r and obtains the values of $U_B || U_A$. Then, it verifies the values of U_A and U_B.

Thus, after the successful completion of this step, both the transmitter and the receiver have the random number which will be used in step 5 to generate the qubits.

Step 3: Creation of a session key

The transmitting user creates a session key (*SK*). This session key is unique for each communication between the users. The first user carries out the following two steps:

- A random number is generated by using a suitable pseudo random function.
- The transmitter signcrypts the session key based on its private key and the public key of the receiver using the RSA-TBOS signcryption scheme and obtains the signcrypted text (c_{SK}).

The aim of this step is to provide a digital signature as well as encryption of the session key for its origin to be validated by the recipient. Moreover, the digital signature guarantees non-repudiation. Furthermore, the use of a cryptographic hash function involved in signcryption can be used to validate the integrity of the data transmitted; that is, it has not been altered.

Step 4: Conversion to Binary

This encrypted session key (c_{SK}) should be converted into binary and then to qubits and finally sent to the corresponding user.

Step 5: Generation of Qubits

Four types of polarization states are used in the generation phase of the quantum bits depending on the pre-shared random number (r) according to Table 8.1.

(1) Vertical represents 0.
(2) Horizontal represents 1.
(3) Down left to upper right '/' represents 1.
(4) Down right to upper left '\' represents 0.

The polarization of a photon can be prepared in any of these states that were mentioned above. Filters exist to distinguish horizontal states from vertical ones. When passing through such a filter, the path of a vertically polarized photon is deflected to the right, while that of a horizontally polarized photon is deflected to the left. In order to distinguish between diagonally polarized photons, one must rotate the filter by 45°. If a photon is passed through a filter with the incorrect orientation—diagonally polarized photon through the non-rotated filter for example—it will be randomly deflected in one of the two directions. In this process, the photon also undergoes a transformation of its polarization state, so that it is impossible to know its orientation before the filter.

The number of bits in the signcrypted session key (c_{SK}) must equal the number of photons. The polarizing state to be used for the polarization of a photon is selected based on the ith bit of the pre-shared random number (r) and the ith bit of the signcrypted session key. If the number of bits in the pre-shared random number is m and the number of bits in the signcrypted session key is n, where m < n, the random number bits are reused; i.e. till the mth bit, the corresponding values will be taken from the random number and for the (m + 1)th bit in the signcrypted session key, the first bit of the random number will be considered and so on.

Table 8.1 Selection of qubit basis

Bit value of (c_{SK})	Bit value of the pre-shared random number	Qubit basis value	Qubit value
0	1	D(diagonal)	\
1	1	D(diagonal)	/
0	0	R(rectilinear)	|
1	0	R(rectilinear)	–

Step 6: Unsigncryption process

Unsigncryption is done on the receiver side based on its private key and the transmitter's public key. The received qubits are measured using the appropriate filters based on the pre-shared random number. After obtaining the binary values, the user converts the resulting binary representation of the signcrypted session key to the equivalent decimal representation. The receiver unsigncrypts the decimal value. This involves verification of the key origin and the decryption of the session key.

After all the previous steps, both users have shared a common session key. This session key is to be used to encrypt the messages to be communicated between those users in future.

Step 7: Exchanging encrypted messages

In this step, a quantum block encryption algorithm is used to assure the confidentiality of the exchanged messages. The established session key in the previous steps is used here. A scheme such as that proposed in [10] can be employed.

8.5 Security Analysis of the Proposed Scheme

In this section, the security properties of the proposed authenticated key distribution scheme are examined

8.5.1 The Intercept/Resend Attack

Let us assume that an eavesdropper (Eve) intercepts the transmitted photons from the transmitter. After a measurement of the photon, Eve resends it to the receiver [11]. This attack cannot break our scheme because when Eve measures the quantum states, she will measure it in wrong bases with probability 0.5. Thus, the receiver will know that the message doesn't come from the original transmitter because the signature part won't be correctly verified most probably. The probability of detecting an eavesdropper in this attack is $(1 - 0.75^n)$, where n is the number of bits in the signcrypted session key.

8.5.2 Beam-Splitting Attack

It is not easy to build a single photon source with current technologies. As a matter of fact, in general, the light pulse referred to as a single photon in the laboratory is not a pure single photon state (i.e., zero, one or multiple photons in the same state). Therefore, the following attack is possible against BB84. First, Eve collects a fraction of the multiple photons by putting a beam-splitter in the path between the transmitter and the receiver. Eve stores the extra photons in a quantum memory until Bob detects the remaining single photon and Alice reveals the encoding basis. Eve can then measure her photons in the correct basis and obtain information on the key without introducing detectable errors [12].

However, this attack is not possible against the proposed protocol. Although, Eve can store the collected photon(s), Eve will not know the quantum state which is being transmitted because Eve doesn't know the random number required for bases synchronization which will never be disclosed in public.

8.6 Discussion

In this section, a comparative study will be provided pointing out the advantages of the proposed scheme over other schemes in literature. The comparison will be held among the following schemes:

1. A quantum authentication protocol using quantum superposition states (Scheme 1).
2. A three party quantum authenticated key distribution protocol using superposition states (Scheme 2).
3. AMNI'09 protocol (Scheme 3).
4. The proposed scheme (Scheme 4).

All these schemes provide:

- Security (or confidentiality)
- Authentication
- Sharing a session key between users

except the first one, which provides authentication only.

Assume a network of n users that need to communicate with each others. Table 8.2 provides the comparison.

Table 8.2 Comparison between the schemes

Schemes	Presence of TC	Information available to TC
1	No TC	No TC
2	TC	TC doesn't know the session key
3	TC	TC knows the session key
4	TC	TC doesn't know the session key

Schemes	No. of rounds in case of two users	No. of long-term keys stored per user
1	3 rounds	(n-1) keys
2	3 rounds per bit and other 4 rounds	1 key
3	2 rounds	1 key
4	3 rounds	2 keys

It is clear from the above table that no third party knows the session key in our protocol as in the first two schemes. However, the proposed protocol is advantageous over the first protocol with regard to the number of long-term keys stored per user and it is advantageous over the second scheme from the viewpoint of the number of rounds required to establish the session key. Finally, the proposed protocol is superior to the third scheme since the session key is only known to the communicating parties with a comparable performance in the number of rounds and storage requirements. Additionally, the proposed protocol ensures integrity of the shared session key.

8.7 Implementation

The proposed protocol has been implemented using Sage software package (www.sagemath.org) under Ubunto operating system to ensure its timeliness and for validation purposes. Sage is a comprehensive open source software package for studying and solving challenging mathematical problems. It covers the basics of mathematics, including calculus and algebra, and a wide range of advanced areas, such as number theory, cryptography, graph theory, and linear algebra.

In the absence of an eavesdropper, it has been validated that both parties share exactly the same key. The presence of an eavesdropper has been simulated by generating a random number (p) in the interval $(0, 1)$ per bit of the signcrypted session key. If p is less than $(1-0.75^{n})$- with n being the number of bits in the signcrypted session key- then the bit is flipped before the unsigncryption process.

The algorithm parameters in one of the sample runs used for validation were as follows: The public key (e), the private key (d) and the modulus (n) of the first user are shown below.

```
   e=7300983133248051083191688582755326987825128503508063194412978025 3
2729771517927318874342756076061224211711651699032618643182305811516490
4015898857437764362479809809728558619465875379885368419804364512607183
3244724865585493157904878754478326674804263705719942836204162367815716
38691833980481740221890581103475 9
   d=9070424659487977418682360326105302321239157279200305714232120285 8
9214837271353905834603043107436096514493768548731422419388295127112392
6126702088799214818691153391098664230202056939063581687090768848343840
5018727501512385921847920713894211223203132332970886578727093562057016
07725521770729201457011706437151 9
   n=3153970747738110327834139729849655725735384686499620649365339528 0
5859847945092613461513320085478575608826307734415943579227687040800697
7046201637134548274874864012369110392717097260778266696594061727721254
0923862436266465237785419536957430810455409509800077482032889712155796
27928164411286909231915920003257 29
```

The public key (e1), the private key (d1) and the modulus (n1) of the second user are given below

```
    e1=3779906161290564671197505319386896171284798225087564214736442317
7299324139458821508436445725655236476388721790624897097382865020897427
6208529947166171021149344888582026438141774219556502918687489068221748
4054692922402595162245856574343279697678749211972313919141248271012100
534720940562555508734860339357481 1
    d1=2716658728151883439313610575396480434472340729572133155301697981
6998278235698426142333513708286238307006168028373008462736960533630390
4273386165835365284027559481835505652250907707968087867100748943176890
5719997966213062946423856283054464163160343001368902943229256659726386
4585470815285343109016537995277870 7

    n1=2886693699230411432443692071623594844233718066829459736449028626

7916959000292956319311810051466844715660550949321478552691369024435912

0573109244400306563678306927435994712985252543284488175302866950157652

7220548177980493233163731425203483818372623640325215500199312752842002

972547821283933662252146363132062 49
```

8.8 Conclusion

In this paper, an optimized key distribution protocol suitable for optical channels has been developed. The protocol involves authenticating the identities of the two communicating parties and confidentiality of the key to be shared between them. A signcryption scheme is used to achieve confidentiality, authenticity and non-repudiation and quantum bits are transmitted over an optical channel to achieve detection of eavesdropping based on the uncertainty principle. The advantages of the use of the signcryption protocol rather than a sign-then-encrypt protocol are that it is computationally more efficient, and saves bandwidth. An advanced quantum block encryption, such as the scheme in [10], is used to achieve the secrecy of messages being transmitted. The session key is known only to the transmitter and the receiver. The trusted center doesn't know the session key. Clearly, the proposed protocol is resistant to various attacks such as the intercept-resend attack and the beam-splitting attack. The use of a public key infrastructure enables reducing the number of keys stored per user which is an important feature in multi-user setting.

References

1. Elboukhari M, Azizi M, Azizi A (2010) Quantum key distribution protocols. Int J Univ Comput Sci I(201):59–67
2. Amutha B, Nivedha V (2009) AMNI'09 protocols. Int J Univ Comput Sci 2:297–303
3. Stallings W (2005) Cryptography and network security: principles and practice, 4th edn. Prentice Hall, New York

4. Bennett CH, Brassard G (2004) Quantum cryptography: public key distribution and coin tossing. In: Proceedings of the International Conference on Computers, Systems and Signal Processing, Bangalore, India
5. Kanamori Y, Moo Yoo S, Gregory D, Sheldon T (2005) On quantum authentication protocols. In: Proceedings of the IEEE Global Tele-communications Conference, GLOBECOM'05, vol 3
6. Li X, Chen L (2007) Quantum authentication protocol using bell states, First International Symposium on Data, Privacy and E-Commerce, IEEE Computer Society, 2007, pp 128–132
7. Reddy KS, Medapati RK (2011) "Three party Quantum Authenticated Key Distribution Protocol Using Superposition States. Int J Comp Appl 2, Rado GT, Suhl H eds Academic Press, New York, pp 1589–1594
8. Malone-Lee J, Mao W (2003) Two birds one stone: signcryption using RSA. In: Topics in cryptology-CT-RSA 2003, Lecture notes in computer science, vol 2612. Springer, Berlin, pp 211–225
9. Abozeid S et al (2012) An authenticated key distribution scheme. In: Proceedings of the World Congress on Engineering and Computer Science, 2012, vol. II, WCECS 2012, 24–26, Oct 2012, San Francisco, USA, pp 983–987
10. Cao Z, Liu L (2010) Improvement of one quantum encryption scheme. IEEE International Conference on Intelligent and Computing Systems (ICIS), vol 1, pp 335–339
11. Bouwmeester D, Ekert A, Zeilinger A (2000) The physics of quantum information. Springer, New York
12. Bennett C, Bessette F, Brassard G, Salvail L, Amolin J (1992) Experimental quantum cryptography. J of Cryptol 5:3–28

Chapter 9
Adaptive Controller Design for Two-Link Flexible Manipulator

Rasheedat Modupe Mahamood

Abstract Flexible Link Manipulator Systems (FLMs) have numerous advantages when compared to the rigid link manipulator such as their light weight, ease of manipulation, low energy consumption, faster manipulation and so on. Controlling such system poses many difficulties due to the distributed nature of the system. In this chapter, direct adaptive control is designed based on a hybrid Proportional Integral Derivative (PID) control system for two-link flexible manipulator. The adaptive control algorithm is simple with less computational load and also very efficient. The control law is tested in Matlab/Simulink environment. The effectiveness of the proposed controller is studied with step input signal and square wave input signal, white noise disturbance, and sine wave disturbance. The results show that the proposed adaptive control law is effective and robust to the disturbances. The results are presented and discussed in detail.

Keywords Adaptive control · Flexible link manipulator · Performance index · PID control · Sine wave disturbance · White noise disturbance

9.1 Introduction

In space exploration, robot is required to have light weight because of the space and weight restriction issues. The benefit of using a lighter weight manipulator when compared to the rigid link manipulator include: less power consumption, higher manipulation speed, they require less material, they use smaller actuator, and they are more maneuverable and transportable [1]. Making the weight of the manipulator to be lighter has resulted in flexibility of the manipulator which makes modeling of such system to become very cumbersome. The dynamic behaviour of

R. M. Mahamood (✉)
Department of Mechanical Engineering, University of Ilorin, Ilorin, Nigeria
e-mail: mahamoodmr@unilorin.edu.ng; mahamoodmr2009@gmail.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_9,

such a system is usually described by partial differential equation which is characterized by infinite dimensional distributed parameter system with non-minimum phase property [2]. This complex model is often truncated to reduce the complexity it will pose on the controller design. That is, by reducing the complexity of the model, it will help to simplify the control system design. Truncating the model will affect the performance of the model based controller in real time operation because; the unmodelled part creates a ripple effect on small error in the control system thereby degrading such system. Against this background, controlling of Flexible Link Manipulation System (FLMS) is very challenging because of the flexible modes. Control of single link flexible manipulator is less challenging when compared to the two-link flexible manipulator because of its simpler structure. This is not so with two-link flexible manipulator because of the complexity of the system dynamics and the highly coupled nature of this dynamics making the control of such a system more challenging.

Different control techniques have been applied to FLMS in the literature [1–5], they include: Proportional Integral Derivative (PID) control, robust control, and adaptive control. Of all these control methods, PID control is the most widely used industrial control [6] because of their simple structure, ease of implementation and they are cheap [7]. PID controller gains are usually tuned to suit certain operating condition and after tuning, these gains are kept at these values until the operating condition changes and are tuned again by an operator. When operating condition changes or there are disturbances on the system, the fixed gain controller will not be able to cope with these changes and steady state error will continue to be present in the system until the gains are re-tuned by the operator. One of the ways of dealing with this steady state error is by making the control gains to be very high, but there is a limit to which the gains can be increased because of the limited actuation power [8]. Because of the shortcomings of the fixed gain PID controller, there is need for a system that will constantly tune the PID gains as and when required without the need for operator to be re-tuning the gains. Hence the need for adaptive controller. Adaptive control schemes are based on a common principle of estimating the structured and the unstructured uncertainties which can degrade the fixed gain controller and compensating for their effect before they degrade the system. The structured uncertainties are the system parameters uncertainties while the unstructured uncertainties are the working environment and external uncertainties on the system. All these disturbances affect the performance of the FLMS when in operation. To maintain the desired performance, for example, to maintain accurate input tracking, there is need to have a controller that is able to account for the rigid motion (tracking), the flexible motion (vibration) and eliminate the effect of operational and system disturbances.

Some of the adaptive control schemes proposed in the literature are of high computational load and the intelligent based ones are also complex and not easy to implement and are also slower in operation. In this chapter, a simple adaptive control law for two-link flexible manipulator is presented. The adaptation algorithm is designed to constantly tune a fixed gain PD controller in the earlier developed hybrid PD-PID controller by [9]. The PD controller regulates the rigid body motion, while the PID controller regulates the fast motion dynamics (vibration). The proposed adaptive control law constantly adjusts the Proportional (P) gain and the Derivative (D) gain of the PD controller in order to constantly reject any disturbance that is capable of degrading the control system while tracking a reference trajectory. Extensive study was performed on the proposed controller through simulation and the results are presented and discussed in detail. The proposed controller is tested through simulation in Matlab/Simulink environment and the performance is compared with that of the PD-PID controller. The robustness of the adaptive control law is demonstrated using step input and square wave input, and also white noise and sine wave disturbances. The dynamic model of the two-link flexible manipulator was developed by De Luca and Siciliano [10] using the Lagrange and the assumed mode method.

The rest of the chapter is organised as follows: In Sect. 9.2 the mathematical model of the planar two-link flexible manipulator is presented. The designing of the control law is presented in Sect. 9.3. Section 9.4 gives a comprehensive simulation results and discussion. The concluding remarks and future work are presented in Sect. 9.5.

9.2 Mathematical Modelling

The mathematical model of the planar two-link flexible manipulator used in this chapter (shown in Fig. 9.1) was developed by De Luca and Siciliano [10] using Lagrange and Assumed mode method.

The links are modelled as Euler–Bernoulli beam with proper clamped-mass boundary conditions. Small elastic deflection is assumed and the motion is restricted to the plane of the rigid motion. The compact closed-form dynamic equation of the arms according to [10] is given by:

$$\boldsymbol{B}(q)\ddot{q} + \boldsymbol{h}(q,\dot{q}) + \boldsymbol{K}(q) = \boldsymbol{\tau} \tag{9.1}$$

$$q = f(\theta, \delta) \tag{9.2}$$

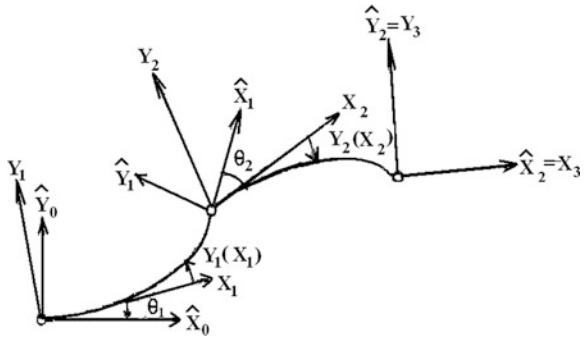


Fig. 9.1 The planar two-link flexible manipulator

where θ is a n-vector of joint coordinates and δ is a m-vector of link deformation coordinates. Let N = n + m, then $\boldsymbol{q}\ (\boldsymbol{\theta}, \boldsymbol{\delta})$ is the N-vector characterising the arms configuration. $\boldsymbol{B}$ is a N × N positive definite symmetric inertial matrix, $\boldsymbol{h}$ is a N-vector containing Coriolis and centrifugal forces and $\boldsymbol{K}$ is a diagonal stiffness matrix. The detailed derivation of the mathematical model can be found in [10] for further reading.

9.3 Controller Design

The adaptive control law is designed to improve the performance of a fixed gain PD controller. The design of the PD-PID controller has been developed in [9] and is briefly explained in Sect. 9.3.1. The PD-PID controller is then extended to incorporate the proposed adaptation law. The detail of the adaptive control law is given in Sect. 9.3.2. In order to compare the overall performance of the PD-PID controller with adaptation and that of the PD-PID controller without adaptation; performance index is determined for the two control schemes and it is presented in Sect. 9..3.3.

9.3.1 PID Controller Design

Figure 9.2 shows the PD-PID control structure developed by [9]. The function of the PD controller is to ensure that the hub follows the reference or the desired trajectories by using the tracking error and the joint velocity in the feedback form to effect the required control. The PID controller on the other hand ensures that the elastic vibrations of the system are controlled simultaneously by using the end-point acceleration also in the feedback form. The PD-PID control input according to [9] is given by:

$$u_{PD_i}(t) = A_{ci}\left(K_{Pi}(\theta_{id}(t) - \theta_i(t)) - K_{vi}\dot{\theta}_i(t)\right) \qquad i = 1, 2 \tag{9.3}$$

where u_{PDi} is PD control input, θ_{id}, θ_i, $\dot{\theta}_i$, A_{ci}, K_{Pi} and K_{vi} are the desired hub angle, actual hub angle, hub velocity, amplifier, proportional and derivative gains respectively.

For PID controller design, the end-point elastic acceleration is used for the control action in a feedback form to control the vibrations in each of the links. Because of the coupling effect, the two links are controlled simultaneously. The control input for the PID according to [9] is also given by:

$$u_{PID_j}(t) = \left(k_{Pj}e(t) + k_{Ij}\int e(t)dt + k_{Dj}(de(t)/dt)\right) \quad j = 1, 2 \tag{9.4}$$

$$e_i(t) = \alpha_{id}(t) - \alpha_i(t) \qquad j = 1,2 \tag{9.5}$$

where u_{PIDj} is the PID controller input, K_{Pj}, K_{Ij}, and k_{Dj} are the proportional, integral and derivative gains. α_{id} (t) and α_i (t) are the desired and actual end-point acceleration.

α_{id} (t) is set to zero since the objective of the PID controller is to have zero acceleration in the system. Total control input τ_i (t) is then given by:

$$\tau_i(t) = u_{PD_i}(t) + u_{PID_i}(t) \qquad i = 1,2 \tag{9.6}$$

Detailed information regarding the design of hybrid PD-PID controller for the two-link flexible manipulator can be found in [9] for further reading.

9.3.2 Adaptive Control Law

The adaptive control law is aimed at constantly tuning the PD controller gains shown in Fig. 9.2 to reject any disturbance on the system in order to eliminate the steady state error. The structure of the proposed adaptive control law was developed by [11] and it is shown in Fig. 9.3.

From the Eq. (9.6), the adaptive control law is given as follows:

$$U_i(t) = \lambda(t)[u_{PDi}(t)] + u_{PIDi}(t) \qquad i = 1,2 \tag{9.7}$$

$$\begin{aligned} U_i(t) =& \lambda_i(t)[k_{pi}(\theta_{id}(t) - \theta_i(t)) - k_{vi}\dot{\theta}_i(t)] \\ &+ \ [k_{Pi} + k_{Ii} + k_{Di}]e_i(t) \qquad i = 1,2 \end{aligned} \tag{9.8}$$

$$\lambda_i(t) = \frac{\Psi_i}{[\theta_{id}(t) - \theta_i(t)]^2 + 1} \qquad i = 1,2 \tag{9.9}$$

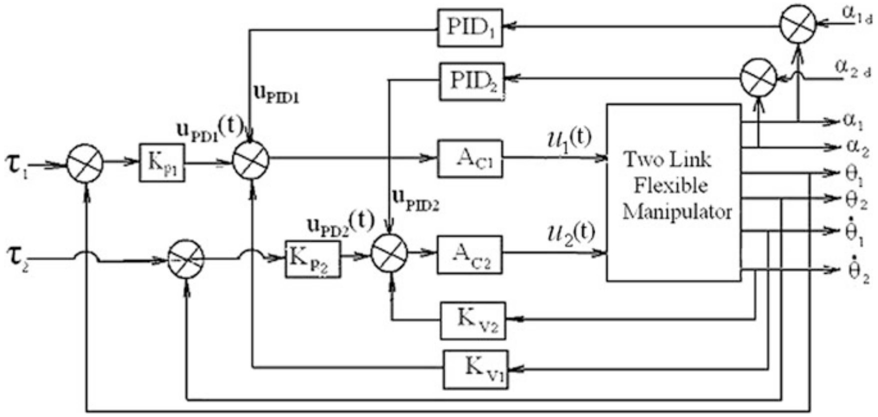


Fig. 9.2 The PD-PID controller structure for the manipulator [9]

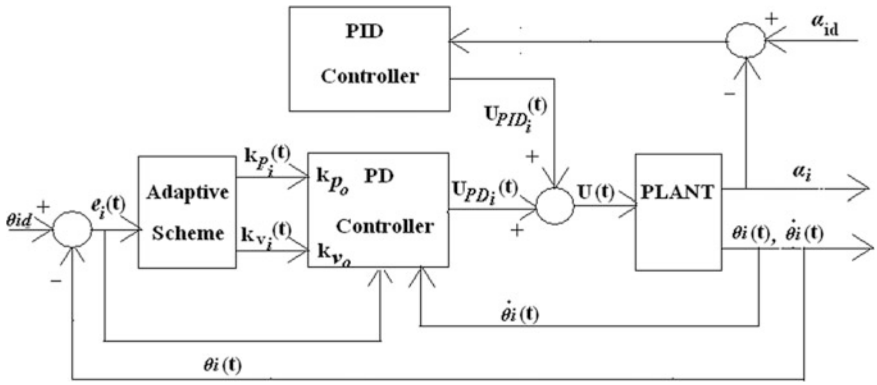


Fig. 9.3 The direct adaptive PD-PID Control architecture

where $u_i\ (t)$ is the total control input, $\lambda_i\ (t)$ is the adaptive parameter that constantly adjusts the PD controller gains, and ψ_i is the adaptive weight gain. Square of the tracking error is used to ensure that the adaptive parameter is stable. That is, the denominator will constantly give a positive result and it will be non zero value so that the controller will not crash.

9.3.3 Performance Index

To be able to compare the performance of the proposed controller with the PD-PID controller, the performance index of the controllers are estimated in terms of input tracking, link velocity, link deflection, input torque and end-point acceleration. The performance index J is given by:

$$J = \frac{1}{t_f}\int_0^{t_f}\left[\sum_{i=1}^{2}\left[\begin{array}{c}\left(\frac{\theta_{id}-\theta_i}{\theta_{i\max}}\right)^2+\left(\frac{\dot{\theta}_{id}-\dot{\theta}_i}{\dot{\theta}_{i\max}}\right)^2+\left(\frac{\delta_{id}-\delta_i}{\delta_{i\max}}\right)^2\\ +\left(\frac{\alpha_{id}-\alpha_i}{\alpha_{i\max}}\right)^2+\left(\frac{u_i}{u_{i\max}}\right)^2\end{array}\right]\right]dt \quad i=1,2 \quad (9.10)$$

where J is the performance index. t_f, θ, $\dot{\theta}$ δ, α and τ are the final simulation time, the hub angle, the link velocity, link deflection, tip acceleration and torque of link i respectively. The subscript imax, id, and i, are the maximum attained value, desired value, and the steady state value of link i respectively.

9.4 Simulation Results and Discussion

The adaptive control law is tested through simulation in Matlab/Simulink environment and the results are presented and discussed in detail this section. The parameters of the planar two-link flexible manipulator used in this study are presented in Table 9.1.

The PD and the PID controller gains were carefully tuned until the desired tracking performance and vibration control was achieved. The adaptive weight gains are also carefully tuned while disturbances are being introduced to ensure that the disturbances are completely rejected and the desired performance is maintained even in the presence of disturbances on the system. The PD, the PID, and the adaptive weight gains are presented in Table 9.2.

The performance of the adaptive control law is studied using step and square wave inputs of 45° and 0.4 rad respectively. To also study the robustness of the proposed control law, white noise and sine wave disturbances were introduced on the system. The responses obtained from the adaptive controller are compared with that of the PD-PID controller and the results are shown in Figs. 9.4, 9.5, 9.6, 9.7, and 9.8. The results are presented and fully discussed in Sects. 9.4.1–9.4.5.

9.4.1 Effect of 45° Step Input on the Performance of the Adaptive Controller

The input tracking performance with the adaptive controller and PD-PID controller are compared and shown in Fig. 9.4a. A smoother tracking without overshoot is

Table 9.1 Two-link flexible manipulator parameters

Symbol	Parameter	Value
$\rho_1 = \rho_2$	Mass density	0.2 kg m^{-3}
$EI_1 = EI_2$	Flexural rigidity	1.0 Nm^2
$l_1 = l_1$	Length	0.51 m
$Jh_1 = Jh_2$	Mass moment of inertia of the hub	0.1 kg m^2
G	Gear ratio	1
$M_1 = m_1$	Mass of the link	0.102 kg
Mp	Mass of pay load	0.102 kg
$J_{o1} = J_{o2}$	Mass moment of inertia of the link about its hub	0.0083 kg m^2
J_p	Mass moment of inertia of end effector	0.0005 kg m^2

Table 9.2 PD-PID and adaptive controller gains [3]

	PD gains		PID gains			Adaptive gain
	K_p	K_v	K_p	K_I	K_d	Ψ
Link 1	1.1	1.1	0.2	0.001	1.5	20
Link 2	0.25	0.42	0.1	0.1	0.5	1

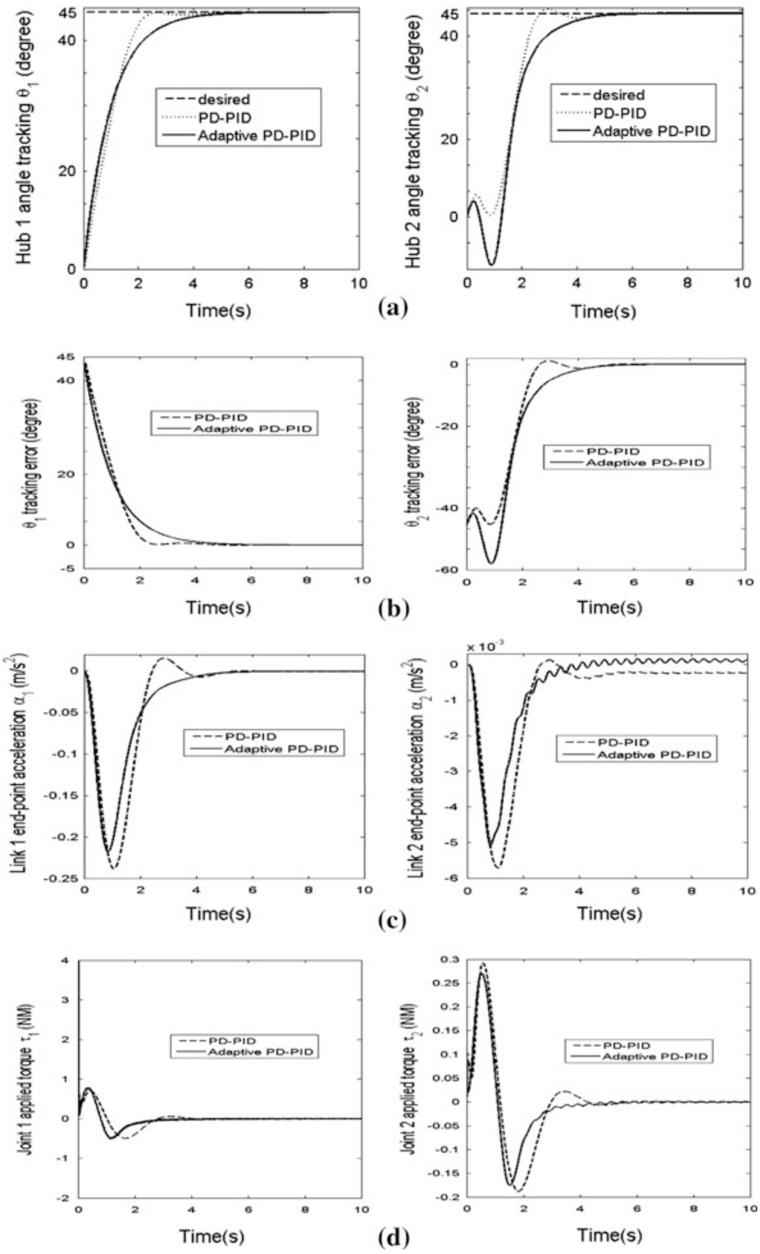


Fig. 9.4 Time history with the 45° step input

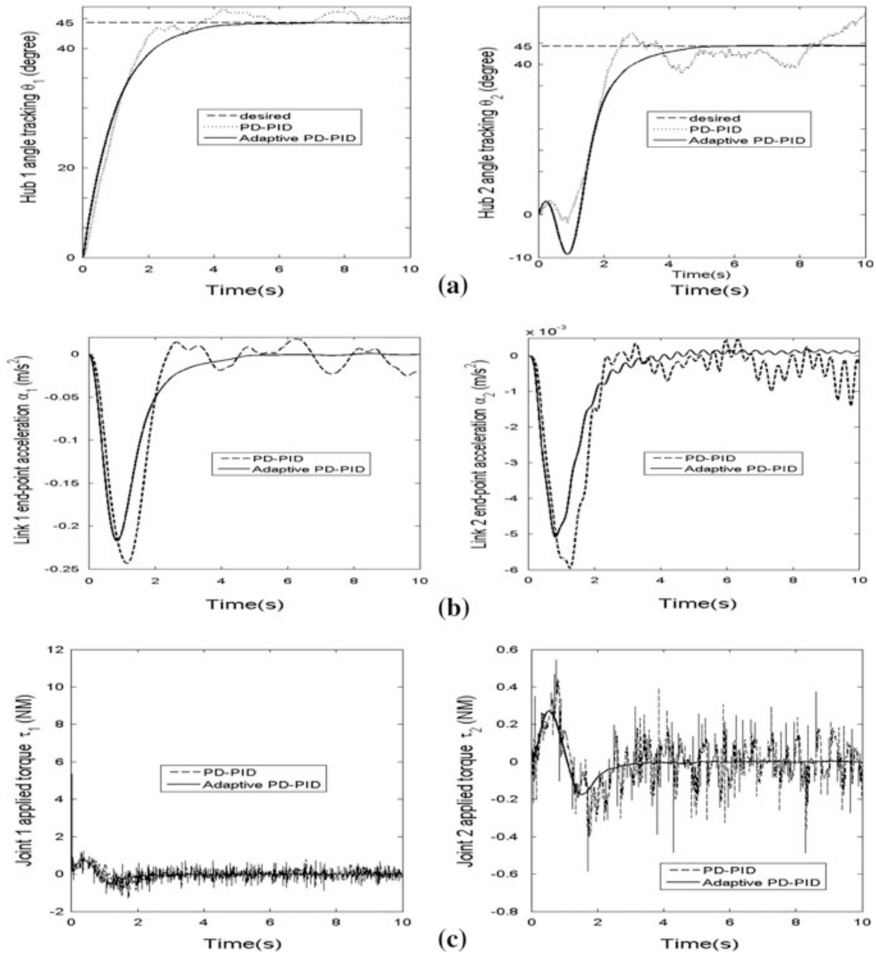


Fig. 9.5 Time history of 45° step input with white noise disturbance

observed with the adaptive controller when compared to the PD-PID controller. The steady state error is reduced with the proposed controller compared to PD-PID controller (see Fig. 9.4b). Figure 9.4c shows the end-points acceleration of the two controllers. It is observed that the amplitude of vibration is reduced with the adaptive controller. It converges closer to zero when compared to that of the PD-PID controller. The total required energy by the adaptive controller is smaller as seen in Fig. 9.4d with the applied torque and it settles down quickly when compared to the PD-PID controller.

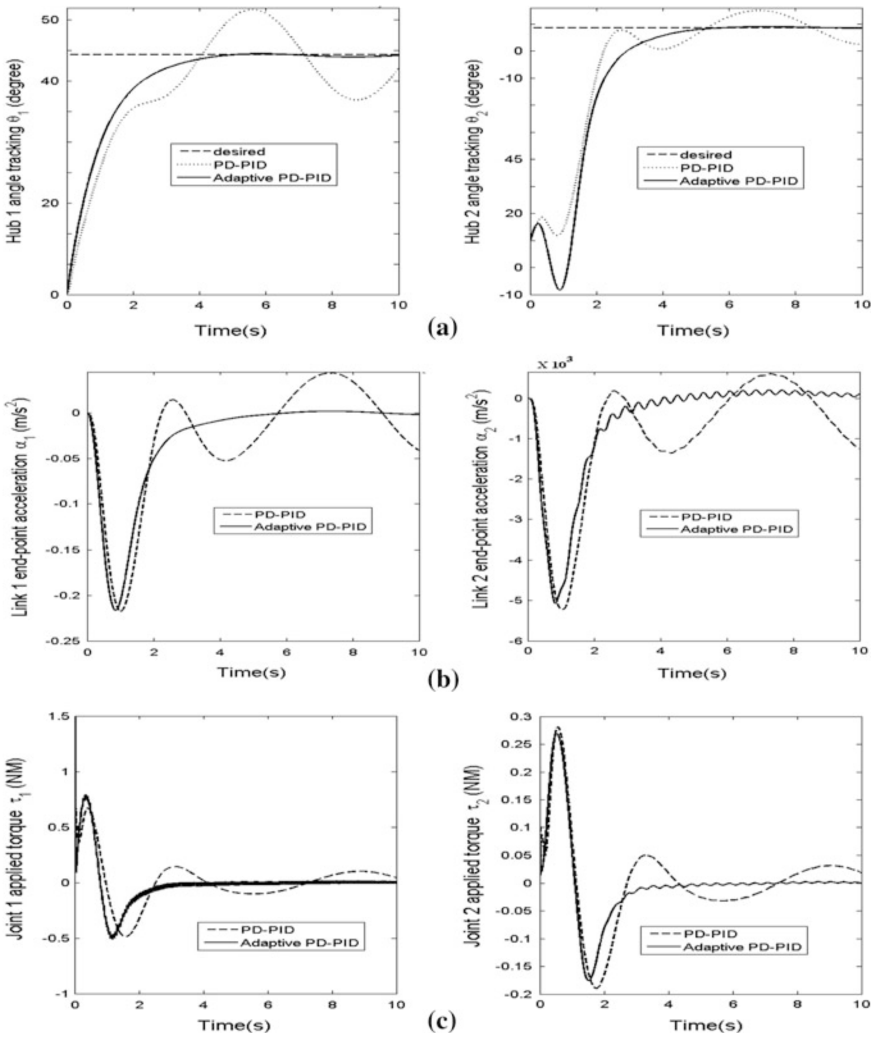


Fig. 9.6 Time history of 45° step input with sine wave disturbance

9.4.2 Effect of White Noise Disturbance on 45° Step Input

To demonstrate the robustness of the adaptive controller, white noise signal is introduced at the first joint and the results are shown in Fig. 9.5.

It is observed that the white noise destabilizes the tracking performance of the PD-PID controller while there is no significant change in the performance of the adaptive controller as shown in Fig. 9.5a, this shows that the adaptive controller is robust to an irregular disturbance like the white noise. Similar result is observed in

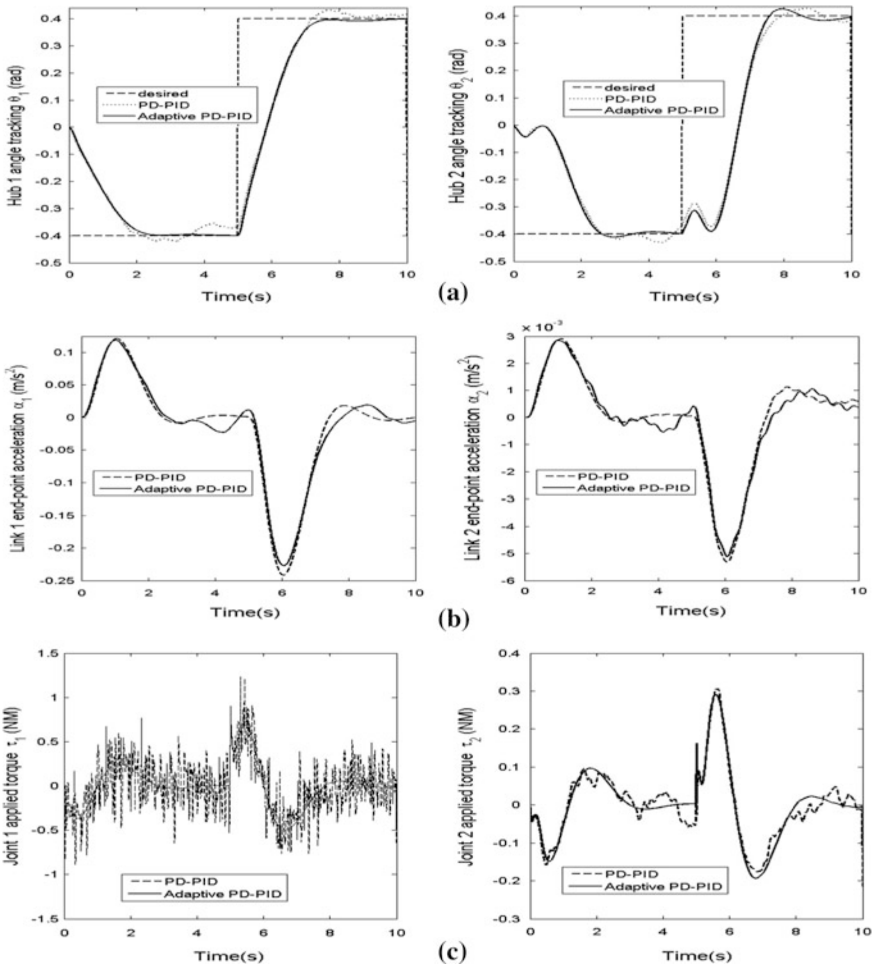


Fig. 9.7 Time history of square input with white noise disturbance

Fig. 9.5b and c, unstable Behaviour were observed in the acceleration and the applied torque respectively with the PD-PID controller, while the adaptive controller shows no change in the applied torque and end point acceleration before the disturbance was introduced and after. The adaptive controller shows no change in its performance compared to the unstable Behaviour observed in the PD-PID controller.

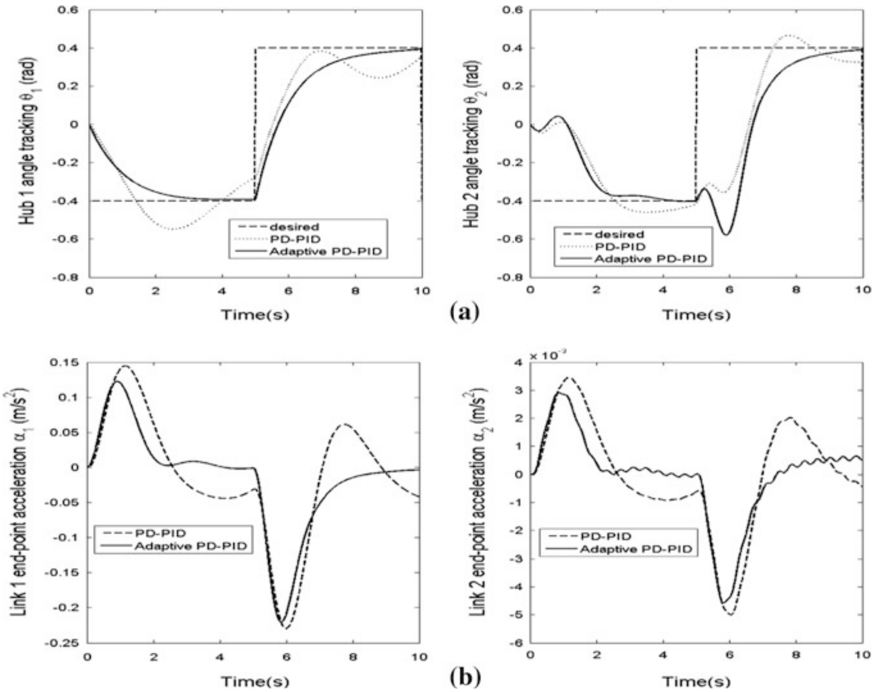


Fig. 9.8 Time history of square wave input with sine wave disturbance

9.4.3 *Effect of Sine Wave Disturbance on 45° Step Input*

The effect of sine wave disturbance, applied to the first joint, on the adaptive controller is also studied (see Fig. 9.6).

Figure 9.6a shows that the tracking performance of the PD-PID controller is unstable. On the other hand the adaptive controller still maintains it performance and it shows no significant change even in the presence of the sine wave disturbance. Figure 9.6b and c also show that the adaptive controller is robust to the disturbance while the PD-PID has become highly unstable, with the end point acceleration and the applied torque becoming very unstable (see Fig. 9.6b and c).

9.4.4 *Effect of the White Noise Disturbance on the Square Wave Input*

To further demonstrate the effectiveness and robustness of the controller, the input was changed to square wave input and white noise was introduced in the first joint and the observed results are presented in Fig. 9.7.

Table 9.3 Performance index

Response	PD-PID		Adaptive PD-PID	
	Link 1	Link 2	Link 1	Link 2
Performance index				
	0.6488		0.6354	

The tracking shown in Fig. 9.7a demonstrated that the adaptive controller is effective. There is no significant change in the performance of the adaptive controller even with the changing input signal to square wave and also in the presence of a noisy disturbance. Tracking is maintained by the adaptive controller while the PD-PID controller failed to track the reference trajectory with unstable and large steady state error. Figure 9.7b and c also show that the performance of the adaptive controller remains unchanged in terms of the end-point acceleration and the applied torque respectively; while that of the PD-PID controller shows unstable Behaviour and chattering in the applied torque.

9.4.5 *Effect of Sine Wave Disturbance on the Square wave Input*

The sine wave disturbance was applied at first joint; it destabilizes the PD-PID controller and the adaptive controller remains unchanged (see Fig. 9.8).

Figure 9.8a shows that the tracking of the adaptive controller is stable while unstable Behaviour is observed with the PD-PID controller. The end-point acceleration shown in Fig. 9.8b was effectively controlled with the adaptive controller whereas, the performance is not stable with the PD-PID controller.

The performance indexes of the two controllers are presented in Table 9.3. The adaptive controller has the overall best performance index as compared to the PD-PID controller which further explains the performance demonstrated by the adaptive controller.

9.5 Conclusion and Future Work

The simplicity of the PID controller makes it use to be more preferred to more advanced controllers unless there is evidence that the PID control cannot handle the requirement at hand. The performance of the PID controllers can be improved using adaptation law in order to be able to cope with unforeseen disturbances on the system. The adaptive control scheme is capable of reducing steady state tracking error resulting from changing input signal or disturbances on the system during operation. The effectiveness of the adaptive control law developed has been demonstrated in this chapter, to improve the performance of the highly non-linear planar two-link flexible manipulator system. An adaptive PD-PID control algorithm has been developed for planar two-link flexible manipulator and the performance has been compared with that of the PD-PID controller without adaptation. The adaptive law is used to constantly tune the PD controller that is used to control the hub angle motion using the actual hub angle and the hub velocity as feedback. The PID controller uses the end-point acceleration in feedback form for vibration suppression. The adaptive control law was incorporated in to the PD-PID controller using the adaptive weight multiply by the reciprocal of the square of the previous hub angle tracking error plus one, to improve the performance of the system. The adaptation is used to tune the P and D gains thereby eliminating the steady state tracking error produced by system disturbances. The proposed control law has been tested through Simulation in Matlab/Simulink environment. The results have demonstrated that a better performance is achieved with the adaptive controller as compared to that obtained by the PD-PID controller without adaptation. The future work will be to carry out experimental validation to ensure the reliability of the proposed controller since this controller will be applied in real time. This will also aid in the transfer of the knowledge from academia to the industry.

References

1. Dwivedy SK, Eberhard P (2006) Dynamic analysis of flexible manipulators. A Lit Rev Mech Mach Theory 41(7):749–777
2. Tokhi MO, Azad AKM (2008) Flexible robot manipulators: modelling, simulation and control. Institution of Engineering and Technology, London
3. Azad AKM (1994) Analysis and design of control mechanisms for flexible manipulator systems. PhD thesis, Department of Automatic Control and Systems Engineering, The University of Sheffield, UK
4. Poerwanto H (1998) Dynamic simulation and control of flexible manipulator systems. PhD thesis, Department of Automatic Control and Systems Engineering, The University of Sheffield, UK
5. Sutton RP, Halikias GD, Plummer AR, Wilson DA (1998) Modelling and H∞ control of a single-link flexible manipulator. Proc Instn Mech Engrs 213(1):85–104
6. Ho MT, Tu YW (2005) PID controller design for a flexible link manipulator. In: Proceedings of the 44th IEEE conference on decision and control and European control conference, pp 6841–6846
7. Ang KH, Chong G, Li Y (2005) PID control system analysis, design, and technology. IEEE Trans Control Syst Technol 13(4):559–576
8. De Luca A, Panzieri S (1994) An iterative scheme for learning gravity compensation in flexible robot arms. Automatica 30(6):993–1002
9. Mahamood MR, Pedro JO (2011) Hybrid PD/PID controller design for two-link flexible manipulators. In: Proceedings of the 8th Asian control conference (ASCC), Kaohsiung, Taiwan, pp 1358–1363
10. De Luca A, Siciliano B (1991) Closed-form dynamic model of planar multilink lightweight robots. IEEE Trans Syst Man Cybern 21(4):826–839
11. Mahamood MR (2012) Direct adaptive hybrid PD-PID controller for two-link flexible robotic manipulator, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science, WCECS, Oct 24–26 2012, San Francisco, USA, pp 1127–1132

Chapter 10
Fuzzy Adaptive Control for a Class of Non-Affine Systems Based on Singular Perturbation Theory

Daoxiang Gao, Dunmin Lu and ZengQi Sun

Abstract A fuzzy adaptive control method is proposed for a class of non-affine nonlinear systems. By combing implicit function theorem and time scale separation, the control input is derived from the solution of a fast dynamical equation. Stability analysis shows that the proposed approach can guarantee the boundedness of the tracking error semi-globally, which can be made arbitrarily small by choosing appropriate design parameters. Tracking performance is illustrated by simulation results.

Keywords Adaptive control · Back-stepping · Dynamic inversion · Fuzzy control · Non-affine systems · Singular perturbation · Time scale separation

10.1 Introduction

Tremendous researches have been made in recent years in the area of controller design for nonlinear systems. Many remarkable results and new design tools, for instance, back-stepping design, fuzzy and neural networks adaptive control methods, were facilitated by advances in geometric nonlinear control theory, in particular, feedback linearization method [9], by which the nonlinear system is transformed into a linear one, then linear control design methods can be applied to

D. Gao (✉) · D. Lu
School of Technology, Beijing Forestry University, Beijing 100083, China
e-mail: dausson@163.com

D. Lu
e-mail: dunminlu@yahoo.com.cn

Z. Sun
State Key Laboratory of Intelligent Technology and system, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
e-mail: szq-dcs@tsinghua.edu.cn

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_10,

acquire the desired performance. Most of these researches are devoted to the control problem of the nonlinear systems in affine form, which are characterized by the control input appearing linearly in the system state equation. For the non-affine systems, the implicit function theorem [1] is commonly used to demonstrate the existence of the optimal solution for the control input, but it does not provide a way to construct such controller to achieve control objective.

Because it is difficult to invert the non-affine nonlinearities to obtain the inverting control input, the fuzzy logic systems (FLs) or neural networks (NNs) are used to be the approximators to approximate the desired feedback control input. In [5, 11, 15, 16], under some assumptions on the original system, several direct adaptive controllers based on NNs or FLs are proposed to deal with the control problem of non-affine system. These approaches use the adaptive controller to approximate the optimal control input directly with a parameter adaptation law designed by the Lyapunov theorem. These direct adaptive control methods are further extended to the output feedback adaptive control in [3, 6, 7, 13]. The main feature of the previous approach is that the uncertainty to be approximated by an adaptive signal contains the adaptive signal itself as a part of uncertainty, which leads to a fixed-point problem. Thus, it needs to involve more restriction on both the input magnitude and the input change. The indirect adaptive control method is concentrated on transforming the original system non-affine in control input to a new system in which the new control input variable appears in an affine form. In [17], FLs are used to approximate the plant model and the control input can be solved by inverting the fuzzy model in an affine form. In [14], the authors use Taylor series expansion to transform the original non-affine system into the affine-like one, then the well-developed adaptive control scheme for affine nonlinear system can be used directly to the non-affine one. However, the indirect adaptive approach has the drawback of the controller singularity problem. In [2], a dynamic feedback adaptive control method is presented by differentiating the original nonlinear state equation once so that the resulting augmented state equation appears linear in the new state variable—the derivative of the control input, which can be used as new control variable. Recently, in [8, 12], the authors propose the control method where the control input is derived from a solution of fast dynamical equation and is shown to stabilize the original non-affine system asymptotically by using Tikihonov theorem directly. By using the fuzzy logic system, in [4], the fuzzy adaptive controller is developed for the unknown non-affine system by the combination of time scale separation and dynamic inversion. The Lyapunov method is used for the stability analysis.

In this chapter, we develop a fuzzy adaptive controller for non-affine systems by using the idea of time scale separation. Unlike [8, 12], the Lyapunov theory is used to show the system stability instead of using Tikhonov theorem directly. The error system dynamics are constructed to facilitate the stability analysis by combing the implicit function theorem and the mean value theorem. we introduce the generalized fuzzy hyperbolic tangent model (GFHM) [18] to be the fuzzy basis function. The properties of the hyperbolic tangent function are exploited to design the adaptive law to guarantee the existence of the solution for control input.

This chapter is organized as follows. Section 10.2 presents a class of non-affine systems that will be considered and some assumptions to facilitate the controller design. In Sect. 10.3, a brief description of GFHM is presented and fuzzy adaptive controller is designed for the unknown non-affine systems Finally, An illustrative example and some conclusion remarks are given in Sects. 10.4 and 10.5.

10.2 Problem Formulation

Consider the following non-affine nonlinear system,

$$\begin{aligned} \dot{x}_i &= x_{i+1} \quad i = 1, \ldots, n-1 \\ \dot{x}_n &= f(x, u) \\ y &= x_1 \end{aligned} \tag{10.1}$$

where $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in R^n$, is the state vector of the system which is assumed available for measurement, $u \in R$ is the scalar control input, $y \in R$ is the system output and the function $f : R^{n+1} \to R$ is a smooth nonlinear function.

The following assumptions are made for system (10.1).

Assumption 1 The map $f : R^{n+1} \to R$, is C^1, $f(0,0) = 0$.

Assumption 2 The inequality $g(x,u) = \partial f(x,u)/\partial u \neq 0$ holds for $(x,u) \in R^{n+1}$. It is implied that $g(x,u)$ is either positive or negative and is bounded away from zero for $(x,u) \in R^{n+1}$. Without losing the generality, we assume that there exist $\overline{g} > \underline{g} > 0$, such that $\overline{g} > g(x,u) > \underline{g} > 0$. This assumption is made for the controllability of the system (10.1) because $g(x,u)$ can be viewed as the control gain of the system (10.1).

Assumption 3 The given desired trajectory $y_d(t)$ and its derivatives up to $(n+1)$-th are bounded.

The aim of this chapter is to find u for system (10.1) such that the system output $y(t)$ tracks a desired trajectory $y_d(t)$ while keeping all the signals of the closed-loop system bounded.

10.3 Fuzzy Adaptive Controller Design

In this section, the GFHM-based fuzzy logic system is used to approximate the unknown function and the adaptive law is designed to ensure the existence of the control input.

10.3.1 Description of GFHM-Based Fuzzy Logic System

The fuzzy logic system performs a mapping from an input vector $x = [x_1, \ldots, x_n]^{\mathrm{T}} \in U \subseteq R^n$ to a scalar output $y \in R$. In this chapter, we use GFHM as fuzzy basis function, which can be characterized by a set of if-then rule as follows [18],

If χ_1 is $F^l_{\chi_1}$ and χ_2 is $F^l_{\chi_2}$ and $\cdots$ and χ_ω is $F^l_{\chi_\omega}$,

Then $y^l = \sum_{m=1}^{\omega} c^l_{F_{\chi m}}$ where, $\chi = [\chi_1, \ldots, \chi_\omega]^{\mathrm{T}}$ is the generalized input vector, derived from the transformation of input vector $x = [x_1, \ldots, x_n]^{\mathrm{T}}$,

$$\begin{array}{lll} \chi_1 = x_1 - d_{11} & \cdots & \chi_{r_1} = x_1 - d_{1r_1} \\ \chi_{r_1+1} = x_2 - d_{21} & \cdots & \chi_{r_1+r_2} = x_2 - d_{2r_2} \\ \vdots & & \vdots \\ \chi_{r_1+\cdots+r_{n-1}+1} = x_n - d_{n1} & \cdots & \chi_{r_1+\cdots+r_n} = x_n - d_{nr_n} \end{array}$$

$\omega = \left(\sum_{j=1}^{n} r_j\right)$ is the total number of the generalized input variables of the fuzzy logic system, r_j are the numbers of the generalized input variables derived from transforming x_j, $(j = 1, \ldots, n)$. d_{ji} $(j = 1, \ldots, n,\ i = 1, \ldots, r_j)$ is the tranformation offset for x_j. F_{χ_m} is the fuzzy sets with respect to χ_m, $m = 1, \ldots, \omega$, which includes only two linguistic expressions, i.e., Positive (P_{χ_m}) and Negative (N_{χ_m}), with respect to which $c_{P_{\chi_m}}$ and $c_{N_{\chi_m}}$ are consequent parameters. The membership function of the fuzzy sets P_{χ_m} and N_{χ_m} for the generalized input variables are depicted as

$$\begin{aligned} \mu_{P_{\chi_m}} &= \exp\left[-(\chi_m - \bar{k}_m)^2/2\right] \\ \mu_{N_{\chi_m}} &= \exp\left[-(\chi_m + \bar{k}_m)^2/2\right] \end{aligned} \tag{10.2}$$

where, $\bar{k}_m$, $m = 1, \ldots, \omega$, is a constant offset. According to these fuzzy rules, the fuzzy logic system with singleton fuzzifier, product inference engine and center-average defuzzifier, is in the following form,

$$y = \theta^{\mathrm{T}} W(x) \tag{10.3}$$

where $\theta = [C_0, C_1]^{\mathrm{T}}$, $C_0 = \sum_{m=1}^{\omega} \frac{c_{P_{\chi_m}} + c_{N_{\chi_m}}}{2}$, $C_1 = \left[\frac{c_{P_{\chi_1}} - c_{N_{\chi_1}}}{2}, \ldots, \frac{c_{P_{\chi_\omega}} - c_{N_{\chi_\omega}}}{2}\right]$, $W(x) = [1, \tanh(\bar{k}_1\chi_1), \ldots, \tanh(\bar{k}_\omega\chi_\omega)]^{\mathrm{T}}$.

It has been proven that GFHM-based fuzzy logic system (10.3) can approximate any continuous function over a compact set $D \subset R^n$ to arbitrary accuracy [18]

$$f(x) = \theta^{*\mathrm{T}} W(x) + \zeta(x) \tag{10.4}$$

where θ^* is the optimal weight parameter, and $\zeta(x)$ is the approximate error. For simplicity, $\zeta(x)$ is denoted by ζ. We assume that there exist optimal weight parameters such that, $|\zeta| \le \zeta_M$ with constant $\zeta_M > 0$ for all $x \in D$. Moreover, θ^* is bounded by $||\theta^*|| \le \theta_M$ on the compact set D, where $\theta_M > 0$ is a constant. Let θ be the estimate of θ^*, and the weight parameter estimation error be $\tilde{\theta} = \theta - \theta^*$.

Remark 1 A property of GFHM is used in this chapter to facilitate our controller design, i.e., $0 \le |\tanh(Z)| \le 1, 0 \le |\partial \tanh(Z)/\partial Z| = |1 - \tanh^2(Z)| \le 1, \forall Z \in R$.

10.3.2 GFHM-Based Fuzzy Controller Design

In this section, we develop a fuzzy controller by use of implicit theorem [1] and singular perturbation theory [10] for the case where the plant model (10.1) is assumed to be unknown.

Let $e = x_1 - y_d$ and the corresponding tracking error vector is $\mathbf{e} = [e, \dot{e}, \ldots, e^{(n-1)}]^{\mathrm{T}}$. We define the filtered tracking error as

$$\xi = [\mathbf{k}^{\mathrm{T}} \ 1]\mathbf{e} \tag{10.5}$$

where $\mathbf{k} = [k_1, \ldots, k_{n-1}]^{\mathrm{T}}$ is determined so that $s^{n-1} + k_{n-1}s^{(n-2)} + \cdots + k_1$ is Hurwitz, i.e., $\mathbf{e} \to 0$ as $\xi \to 0$. From (10.1), the following error dynamic is immediate

$$e^{(n)} = f(x, u) - y_d^{(n)} \tag{10.6}$$

Adding and subtracting bu, we can rewrite (10.6) as

$$e^{(n)} = f(x, u) - bu + bu - y_d^{(n)} \tag{10.7}$$

where $b > 0$ is a design constant, which will be specified later. The term bu is used to ensure the existence of the control input for the tracking problem. The fuzzy logic system is used to approximate the unknown function $f_b(x, u) = f(x, u) - bu$

$$f_b(x, u) = \theta^{*\mathrm{T}} W(x, u) + \zeta \tag{10.8}$$

then, we have

$$e^{(n)} = \theta^{\mathrm{T}} W(x, u) - \tilde{\theta}^{\mathrm{T}} W(x, u) + \zeta + bu - y_d^{(n)} \tag{10.9}$$

where $\theta = [\theta_x^{\mathrm{T}}, \theta_u^{\mathrm{T}}]^{\mathrm{T}} \in R^{\omega}$, $W(x, u) = [W(x)^{\mathrm{T}}, W(u)^{\mathrm{T}}]^{\mathrm{T}} \in R^{\omega}$, $\omega =$ d $\left(\sum_{i=1}^{n} r_i + r_u\right)$ is the total number of the generalized input variables of the fuzzy system. r_i, $i = 1, \cdots, n$ and r_u are the numbers of the generalized inputs by transforming x_i and u. ${\theta_x}^{\mathrm{T}}$ and ${\theta_u}^{\mathrm{T}}$ are the estimation parameters with respect to

$W(x)$ and $W(u)$. ζ is the approximate error. Let $v = K\xi + [0\ \mathbf{k}^{\mathrm{T}}]\mathbf{e} - y_d^{(n)}$, $K > 0$, then the error dynamic is as the following

$$\dot{\xi} = -K\xi + [\theta^{\mathrm{T}} W(x,u) + bu + v - \tilde{\theta}^{\mathrm{T}} W(x,u) + \zeta] \tag{10.10}$$

We design the following fast dynamic system to obtain the control input u,

$$\epsilon \dot{u} = -\theta^{\mathrm{T}} W(x,u) - bu - v \tag{10.11}$$

If we choose $\varepsilon = 0$, (10.11) is reduced to an algebraic equation

$$0 = -\theta^{\mathrm{T}} W(x,u^*) - bu^* - v \tag{10.12}$$

where u^* is the desired control input of u. The existence of u^* is demonstrated by implicit theorem [1]. Because ξ, $[0\ \mathbf{k}^{\mathrm{T}}]\mathbf{e}$ and $y_d^{(n)}$ are not the functions of u, we have

$$\begin{aligned} \hat{g}(x,u) &= \partial[\theta^{\mathrm{T}} W(x,u) + v + bu]/\partial u \\ &= \theta_u^{\mathrm{T}} \partial W(u)/\partial u + b \end{aligned} \tag{10.13}$$

In order to guarantee the existence of the solution for the control input in (10.11) and (10.12), it should be ensured that $\hat{g}(x,u) \neq 0$. To be consistent with the assumption 2 for original system (10.1), let $\hat{g}(x,u) > 0$ and $b > 0$. We should design an adaptive law and select an appropriate constant b_1 such that $\hat{g}(x,u) > b_1 > 0$ is bounded away from zero. From remark 1, the following inequality holds,

$$\begin{aligned} |\hat{g}(x,u)| &\geq |b| - ||\theta_u^{\mathrm{T}}|| * ||\partial\vartheta(\upsilon)/\partial\upsilon|| \\ &\geq b - ||\theta_u^{\mathrm{T}}|| \\ &> b_1 \end{aligned} \tag{10.14}$$

Thus, the adaptive law should ensure $||\theta_u^{\mathrm{T}}|| < b - b_1$. We can get $b_1 < \hat{g}(x,u) < 2b - b_1$. Let $\hat{f}_b(x,u) = \theta^{\mathrm{T}} W(x,u) + bu$. Using mean value theorem, there exists $0 < \lambda < 1$, such that

$$\hat{f}_b(x,u) = \hat{f}_b(x,u^*) + \hat{g}_\lambda (u - u^*) \tag{10.15}$$

where $\hat{g}_\lambda = \hat{g}_\lambda(x, u_\lambda)$, $u_\lambda = \lambda u + (1-\lambda)u^*$. Note that $2b - b_1 > \hat{g}_\lambda > b_1 > 0$.

Let $\eta = u - u^*$. From (10.10)–(10.12) and (10.15), we obtain the following error dynamics (a general form of singular perturbed system [10]).

$$\begin{cases} \dot{\xi} &= -K\xi + \hat{g}_\lambda \eta - \tilde{\theta}^{\mathrm{T}} W(x,u) + \zeta \\ \varepsilon\dot{\eta} &= -\hat{g}_\lambda \eta - \varepsilon \dot{u}^* \end{cases} \tag{10.16}$$

The adaptive law is

$$\dot{\theta}_x = \Gamma_1[W(x)\xi - \delta_1\theta_x]$$

$$\dot{\theta}_u = \begin{cases} \Gamma_2 W(u)\xi & \text{if } ||\theta_u|| < b - b_1 \\ & \text{or } ||\theta_u|| = b - b_1 \\ & \text{and } \xi\theta_u^T W(u) \le 0 \\ \Gamma_2 W(u)\xi - \Gamma_2 \frac{\xi\theta_u^T W(u)}{||\theta_u||^2}\theta_u & \text{if } ||\theta_u|| = b - b_1 \\ & \text{and } \xi\theta_u^T W(u) > 0 \end{cases} \tag{10.17}$$

where Γ_1 and Γ_2 are positive defined diagonal matrices, $\delta_1 > 0$ is a constant.

Consider the Lyapunov function,

$$V = 1/2\left(\xi^2 + \eta^2 + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\tilde{\theta}\right) \tag{10.18}$$

where $\Gamma = \mathrm{diag}(\Gamma_1, \Gamma_2)$, we have the following theorem.

Theorem 1 Consider the system (10.1) regulated by the control law in (10.11). Suppose that the Lyapunov function (10.18) is bounded by a given positive constant p for all initial conditions, and that the estimation parameters are updated according to (10.17). Then, all the closed loop signals are semi-globally uniformly bounded, and the tracking errorr is attracted to a neighborhood of the origin, whose size can be adjusted by control parameters.

Proof Firstly, we should give the upper bound of $\dot{u}^*$. Differentiate the right hand side of (10.12) and after some simple manipulations, we have

$$\begin{aligned} \dot{u}^* = & -\left[b + \theta_u^{\mathrm{T}}\frac{\partial W(u^*)}{\partial u^*}\right]^{-1}\left[\theta_x^{\mathrm{T}}\frac{\partial W(x)}{\partial(x_1,\ldots,x_n)}(\dot{x}_1,\ldots,\dot{x}_n)^{\mathrm{T}}\right. \\ & \left. + K\dot{\xi} + [0\mathrm{k}^{\mathrm{T}}]\dot{\mathrm{e}} - y_d^{(n+1)}\right] \end{aligned} \tag{10.19}$$

By induction, there exists a continuous function $B(y_d,\ldots,y_d^{(n+1)},\tilde{\theta},\xi,\eta)$, such that,

$$\dot{u}^* = B(y_d,\ldots,y_d^{(n+1)},\tilde{\theta},\xi,\eta) \tag{10.20}$$

Define the compact sets, $D_0 := \{y_d, \dot{y}_d, \ldots, y_d^{(n+1)} | y_d^2 + \dot{y}_d^2 \cdots + \left(y_d^{(n+1)}\right)^2 \le B_0\}$, $D_1 := \left\{\xi^2 + \eta^2 + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\tilde{\theta} \le 2p\right\}$ for $p > 0$. Clearly, $D_0 \times D_1$ is compact in $R^{n+3+\omega}$, where ω is the dimension of $\tilde{\theta}$. Therefore, B has a maximum M on $D_0 \times D_1$.

The derivative of the Lyapunov function is

$$\begin{aligned} \dot{V} &= \xi\dot{\xi} + \eta\dot{\eta} + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\dot{\theta} \\ &= -K\xi^2 + g_\lambda\xi\eta - \tilde{\theta}^{\mathrm{T}}W(x,u)\xi + \xi\zeta - \epsilon^{-1}g_\lambda\eta^2 - \eta\dot{u}^* + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\dot{\theta} \end{aligned} \tag{10.21}$$

Let $\underline{\beta} \leq \hat{g}_\lambda \leq \overline{\beta}$, where $\overline{\beta} = 2b - b_1$, $\underline{\beta} = b_1$.

Using the facts,

$$\xi^2 + (1/4)\eta^2 \geq \xi\eta$$
$$\xi^2 + (1/4)\zeta^2 \geq \xi\zeta$$

we have

$$\begin{aligned} \dot{V} \leq &(-K + 1 + \overline{\beta})\xi^2 + (1/4)\overline{\beta}\eta^2 + (1/4)\zeta^2 \\ &- \epsilon^{-1}\underline{\beta}\eta^2 + |\eta B| + \tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\dot{\theta} - \tilde{\theta}^{\mathrm{T}}W(x,u)\xi \end{aligned} \tag{10.22}$$

Using (10.17), we can deduce the following inequality,

$$\tilde{\theta}_u^{\mathrm{T}}(\Gamma_2^{-1}\dot{\theta}_u - \xi W(u)) \leq 0$$

Substitute the adaptive law (10.17) into (10.22),

$$\dot{V} \leq (-K + 1 + \overline{\beta})\xi^2 + (1/4)\overline{\beta}\eta^2 + (1/4\zeta^2) - \epsilon^{-1}\underline{\beta}\eta^2 + |\eta M| - \delta_1\tilde{\theta}_x^{\mathrm{T}}\theta_x \tag{10.23}$$

Choose $K = 1 + \overline{\beta} + \gamma_0$ and $\epsilon^{-1} = (1/\underline{\beta})[(1/4)\overline{\beta} + M^2/(2\rho) + \gamma_0]$, where γ_0 and ϱ are positive constants. Using $2\tilde{\theta}_x^{\mathrm{T}}\tilde{\theta}_x \geq ||\tilde{\theta}_x||^2 - ||\theta_x^*||^2$, we obtain,

$$\begin{aligned} \dot{V} &\leq \left[-\gamma_0(\xi^2 + \eta^2) + \rho/2 + (1/4)\zeta^2\right] - (1/2)\delta\Big(||\tilde{\theta}_x||^2 - ||\theta_x^*||^2\Big) \\ &\leq \left[-\gamma_0(\xi^2 + \eta^2) - \ \mathrm{d}\ \frac{\delta}{2\lambda_{\max}(\Gamma^{-1})}\tilde{\theta}^{\mathrm{T}}\Gamma^{-1}\tilde{\theta}\right] + \rho/2 + (1/4)\zeta^2 + (\delta/2)||\theta^*||^2 \end{aligned} \tag{10.24}$$

Let $e = (1/4)\zeta^2 + (\delta/2)||\theta^*||^2$, because $|\zeta| < \zeta_M$ and $||\theta^*|| < \theta_M$, then we get $e \leq (1/4)\zeta_M^2 + (\delta/2)||\theta_M||^2 = e_M$. Choose $\gamma = \min\big(\gamma_0, \delta/\left[2\lambda_{\max}(\Gamma^{-1})\right]\big)$,

$$\dot{V} \leq -2\gamma V + (\rho/2 + e_M) \tag{10.25}$$

Let $\gamma > (\rho/2 + e_M)/(2p)$, then $\dot{V} < 0$ on $V(t) = p$. Let $L = (\rho/2 + e_M)/(2\gamma)$ and for all $t \geq 0$, the solution of inequality (10.24) is

$$0 \leq V(t) \leq L + [V(0) - L]\exp(-2\gamma t) \tag{10.26}$$

It means that $V(t)$ is eventually bounded by L. Thus, ξ, η are uniformly bounded. By choosing appropriate value K, ϵ, the quantity L can be made arbitrarily small. Because ξ is bounded, from (10.5), it follows that $\mathbf{e} = \left[e, \dot{e}, \ldots, e^{(n-1)}\right]^{\mathrm{T}}$ is bounded. Then, all the signals in the closed loop system are bounded.

It is clear that increasing the values of K, reducing the value of $\lambda_{\max}(\Gamma^{-1})$ and ϵ, i.e., increasing the value of γ will result in a better tracking performance, but lead to a high gain control scheme. Decreasing δ_1 will help to reduce e, however, a very small δ_1 may not be enough to prevent the fuzzy weight estimates from drifting to very large values.

10.4 Numerical Simulation

In this section, the following example are considered to illustrate the proposed control method.

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_1^3 + x_1 x_2 + u^3/3 + 0.2(1 + x_2^2)u \\ y &= x_1 \end{aligned} \tag{10.27}$$

where, x_1 and x_2 are state variables, u is control input, y is the output. Clearly, system (10.27) is in non-affine form (10.1) and satisfies assumption 1 and 2. The control objective is to find a control input u such that all the signals in the closed-loop system remain bounded and the output y can track the desired trajectory $y_d = 0.75[\sin(t) + \cos(0.5t)]$.

For the unknown system, we design the GFHM-based fuzzy adaptive controller in the following procedure.

We use $[x_1, x_2, u]^{\mathrm{T}}$ as the input vector of the fuzzy logic system. The generalized input vector χ can be obtained by transforming the input variables x_1, x_2, u. In simulation, every input variable is transformed for three times. For x_1 and x_2, the transformation offsets are $d = [-0.5, 0, 0.5]$ and the constant offset is $\overline{k} = 0.8$. For u, the transformation offsets are $d = [-1.5, 0, 1.5]$ and the constant offset is $\overline{k} = 2$. Two membership functions depicted as (10.2) are chosen for each generalized input variable (see e.g., Figs. 10.1 and 10.2). Choose the following design parameters, $k = 2$, $\xi = \dot{e} + ke$, $K = 5$, $\epsilon = 0.02$, $\Gamma_1 = \Gamma_2 =$ diag $\{10\}$, $\delta_1 = 0.1$, $b_1 = 0.2$. The fuzzy adaptive controller is

$$\epsilon \dot{u} = -\theta^{\mathrm{T}} W(x, u) - bu - K\xi - k\dot{e} + \dot{y}_d, \;\; u = \int \dot{u} \mathrm{d}t \tag{10.28}$$

Remark 2 The choice of b is critical for the closed-loop stability and the tracking performance. On the one hand, too small value of b cannot guarantee the existence of solution for u in (10.11) because $b + \partial[\theta^{\mathrm{T}} W(x, u)]/\partial u = 0$ may occur and it leads to the controller singularity, on the other hand, the signal of $f(x, u)$ is submerged by the signal of bu if too large value of b is chosen [see (10.8)] and it may degrade the tracking performance because the actual dynamic of system (10.1) cannot be updated correctly by the fuzzy logic system through on-line learning.

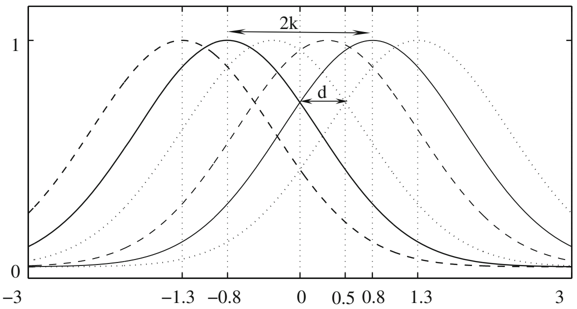


Fig. 10.1 Membership functions of the generalized input variables for x_1 (x_2). *Dashed line* $x_1 - 0.5$; *Solid line* x_1; *Dotted line* $x_1 + 0.5$

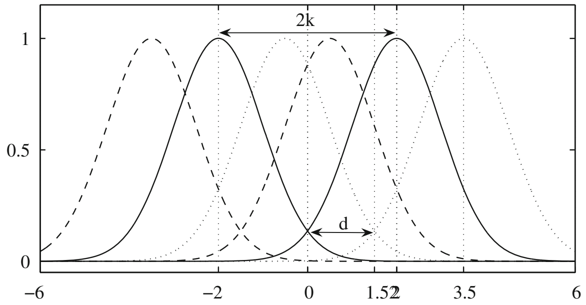


Fig. 10.2 Membership functions of the generalized input variables for u. *Dashed line* $u - 1.5$; *Solid line* u; *Dotted line* $u + 1.5$

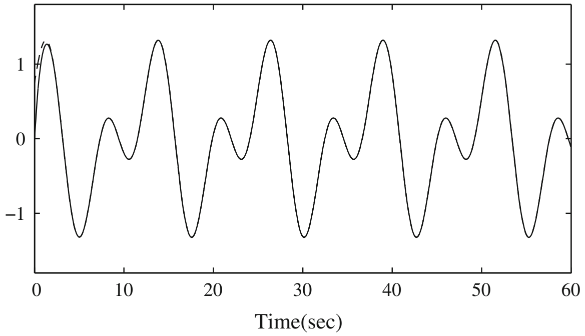


Fig. 10.3 Actual (*solid*) and desired (*dotted*) output ($b = 1$)

According to assumption 2, we choose $\underline{g} < b_1 < b < (\underline{g} + \overline{g})/2$, such that, $\overline{g} > \hat{g}(x, u) > \underline{g} > 0$ holds.

To show how the choice of b affects the control performance of the system, We choose $b = 1, 2$ for the simulation and Figs. 10.3–10.6 show the simulation results with fuzzy adaptive controller. In Fig. 10.5, it can be seen that the controller with $b = 1$ has a better tracking performance and too large value $b = 2$ degrades the performance but still guarantees the system stability. Figures 10.4 and 10.6 show the boundedness of control input and the fuzzy estimation parameters for different choices of b.

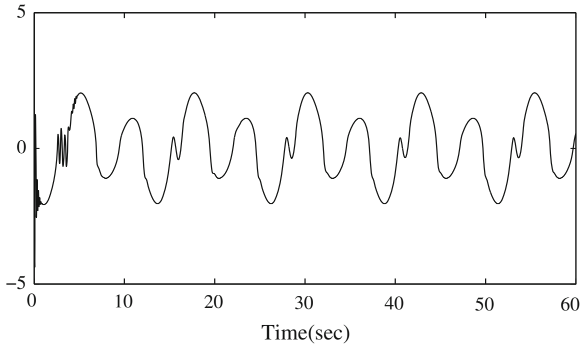


Fig. 10.4 Adaptive control input ($b = 1$)

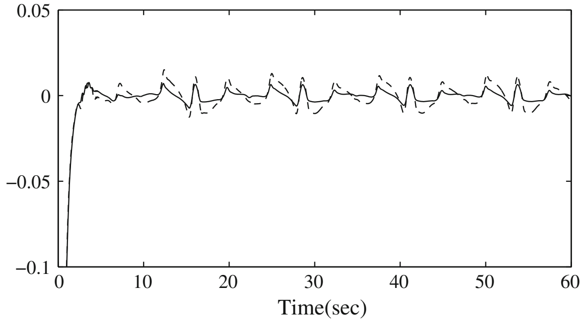


Fig. 10.5 Tracking errors (*solid line* b = 1; *dotted line* b = 2)

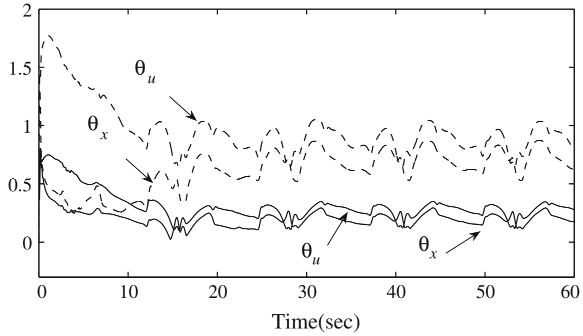


Fig. 10.6 L_2 norms of θ_x, θ_u. *solid line* $b = 1$; *Dotted line* $b = 2$

10.5 Conclusion

In this chapter, time scale separation based fuzzy adaptive control method is developed for a class of non-affine nonlinear systems. The implicit function theorem is used to demonstrate the existence of the optimal control input for the non-affine system, which is approximated by the solution of a fast dynamical equation.

The error system dynamics are constructed by combining implicit function theorem and the mean value theorem. Stability analysis based on Lyapunov theory shows that the developed control scheme achieves semi-globally uniform boundedness of all the signals in the closed-loop, and the bounded errors can be made arbitrarily small by choosing appropriate design parameters.

References

1. Apostol TM (1974) Mathematical analysis. Addison-Wesley, Reading, Mass
2. Bošković JD, Chen LJ, Mehra RK (2001) Adaptive tracking control of a class of non-affine plants using dynamic feedback. In: Processdings of the 2001 American control conference, IEEE Press, Arlington, pp 2450–2455
3. Calise AJ, Hovakimyan N, Idan M (2001) Adaptive output feedback control of nonlinear systems using neural networks. Automatica 37(8):1201–1211
4. Gao DX, Lu DM, Sun ZQ (2012) Fuzzy adaptive control for a class of non-affine systems via time scale separation. Lecture Notes in engineering and computer science. In: Proceedings of The World congress on engineering and computer science, WCECS, 24–26 Oct 2012, USA, pp 1193–1197
5. Ge SS, Zhang J (2003) Neural-network control of nonaffine nonlinear system with zero dynamics by state and output feedback. IEEE Trans Neural Netw 14(4):900–918
6. Hovakimyan N, Nardi F, Calise AJ (2002) A novel error observer-based adaptive output feedback approach for control of uncertain systems. IEEE Trans Automa Control 47(8):1310–1314
7. Hovakimyan N, Nardi F, Calise A, Kim N (2002) Adaptive output feedback control of uncertain nonlinear systems using single-hidden-layer neural networks. IEEE Trans Neural Netw 13(6):1420–1431
8. Hovakimyan N, Lavretsky E, Sasane A (2005) Dynamic inversion for nonaffine-in-control systems via time-scale separation: Part I. In: Proceedings of the 2005 American control conference, IEEE Press, Portland, pp 3542–3547
9. Isidori A (1989) Nonlinear control systems. Springer, Berlin
10. Khalil HK (1996) Nonlinear systems. Prentice- Hall, Upper Saddle River
11. Labiod S, Guerra TM (2007) Adaptive fuzzy control of a class of SISO nonaffine nonlinear systems. Fuzzy Sets Syst 158(10):1126–1137
12. Lavretsky E, Hovakimyan N (2005) Adaptive dynamic inversion for nonaffine-in-control systems via time-scale separation: Part II. In: Proceedings of the 2005 American control conference, IEEE Press, Portland, pp 3548–3553
13. Park JH, Kim SH (2004) Direct adaptive output-feedback fuzzy controller for nonaffine nonlinear system. IEE Proc Control Theory Appl 151(1):65–72
14. Park JY, Park GT (2003) Robust adaptive fuzzy controller for nonaffine nonlinear systems with dynamic rule activation. Int J Robust Nonlinear Control 13(2):117–139
15. Park JH, Park GT, Kim SH, Moon CJ (2005) Direct adaptive self-structuring fuzzy controller for nonaffine nonlinear system. Fuzzy Sets Syst 153(3):429–445
16. Yang BJ, Calise AJ (2007) Adaptive control of a class of nonaffine systems using neural networks. IEEE Trans Neural Netw 18(4):1149–1159
17. Yoon PS, Park JH, Park GT (2002) Adaptive fuzzy control of nonaffine nonlinear systems using Takagi-Sugeno fuzzy models. In: Procedings of IEEE international conference on fuzzy systems, IEEE Press, Hawaii, pp 642–645
18. Zhang HG, Quan YB (2001) Modeling, identification, and control of a class of nonlinear systems. IEEE Trans Fuzzy Syst 9(2):349–354

Chapter 11
Using Multi-Agent Systems for Hardware Upgrade Advice in Smart Grid Simulations

Ala Shaabana, Sami Syed, Ziad Kobti and Kemal Tepe

Abstract The environmental impact of the petroleum-based infrastructure has led to renewed interest in electrical transport infrastructures during the past few decades. However, the impact of plug-in hybrid electric vehicles (PHEV) on electrical distribution and generation systems is not yet fully understood. This poses challenges to distribution and generation companies on how to retool their distribution and generation systems to meet and supply increased future demand. Furthermore, having unpredictable and uncontrollable generation patterns of renewable energy sources in the grid makes it even harder to manage supply and demand in the grid. The ultimate goal is to provide a tool for engineers to further understand and evaluate potential grid infrastructures under different operating conditions. With this simulator, the grid can be evaluated with different hardware and operating conditions to maximize resources. As such, utilities and generation companies can evaluate and test different strategies to upgrade the infrastructure to improve reliability and generation capacity to effectively meet demand.

Keywords Artificial intelligence · Energy conservation · Multi-agent system · Plug-in hybrid electric vehicle · Renewable energy · Simulation · Smart grid

A. Shaabana (✉) · Z. Kobti
School of Computer Science, University of Windsor,
401 Sunset Avenue, Windsor, Canada
e-mail: shaaban@uwindsor.ca

Z. Kobti
e-mail: kobti@uwindsor.ca

S. Syed · K. Tepe
Department of Electrical and Computer Engineering, University of Windsor,
401 Sunset Avenue, Windsor, Canada
e-mail: sami@uwindsor.ca

K. Tepe
e-mail: ktepe@uwindsor.ca

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_11,

11.1 Introduction

Electric vehicles first surfaced in the mid-19th century, when electricity was among the preferred methods of motor vehicle propulsion, providing a level of comfort and ease of operation that could not be achieved by their gasoline counterparts of the time. Electric power has remained commonplace in other vehicle types such as trains and smaller vehicles even though the internal combustion engine is now the dominant propulsion method for motor vehicles. During the last few decades, the environmental impact of the petroleum-based transportation infrastructure has led to renewed interest in an electric transportation infrastructure.

Traditionally, electrical distribution systems (i.e., grids) are designed to provide energy from large generators to the customers. However, environmental concerns such as global warming and pollution force us to utilize small scale Renewable Generation (RG) systems as well as the electrification of the transportation systems, such as plug-in-hybrid electrical vehicles (PHEV). These new generation techniques and increased demand by PHEVs pose great challenges to the existing grid infrastructure since the energy flow can reverse direction from the customer side to grid. Demand by PHEVs can cause grids to collapse, interestingly PHEV batteries can be utilized as storage and supply electricity back to grid to shave off peak demand. This conflicting and challenging paradigm shift requires new design and control mechanisms in the grid. Recognizing this, an initiative called Smart Grid (SG) has been created to better integrate automation, monitoring and managing all the entities with smart devices and communication infrastructure. SG will eventually allow the management of new distributed RG such as solar panels and wind turbines, and innovative power consumers such as PHEV. One of the benefits of SG is to curb infrastructure over-loads and temporary energy shortages, called imbalance costs. Imbalance occurs due to unpredictable changes in production and consumption, hence an imbalance cost is the extra expense for compensating this temporary imbalance between supply and demand.

Energy imbalance is usually curbed with voltage regulation, and when demand increases the voltage drops, hence the imbalance. However, voltage regulation may have negligible effect on power demand, for example electric heaters on a lower voltage simply run longer to deliver the same amount of heat, while devices with electrical motors generate a lower magnetic field and may work longer to complete the same mechanical work. The study done in 2010 by the United State's Department of Energy's Pacific Northwest National Laboratory (PNNL) by Schneider et al. [1] highlights difficulties of regulating supply and demand with voltage regulation. In recent years, SG has been investigated to effectively match supply and demand. One of such investigations is the work done by James et al. [2], which suggests an adaptive, intelligent agent-based software system that can provide real-time, two-way communication and decision making between Distributed Energy Resources (DERs) as system nodes. James et al. advocate that the proposed software also enables the ability to create on-the-fly aggregate blocks of capacity for presentation to the energy markets. This and the work done by Schneider et al. suggest building solutions on top of existing grid systems, which may not solve the problem permanently in certain cases.

Inspired by the work done in the SG field by Schneider et al. and James et al., our objective in this work is to analyse, design, implement and evaluate an approach which models and simulates a dynamic infrastructure of a city wide SG. This model and simulator is designed to provide an SG scenario as close to a real-world scenario as possible using individual agents. The proposed system suggests gradual, hierarchical upgrades to the system as various parameters within the simulation, such as the number of PHEVs on the system. The suggested upgrades change and adapt in cost and power output as time and populations change.

The rest of this article is structured as follows. Section 11.2 presents the general background of our work. Section 11.3 discusses which technologies we have used in our implementation. Section 11.4 will discuss the simulation overview, including placement and power consumption calculation algorithms. Section 11.5 will discuss the correlations between agents as well as the mechanisms in which the hierarchical upgrades are suggested. Finally, Sect. 11.6 will discuss the results generated by the simulation and the conclusion.

11.2 Related Work

Dimeas et al. [3] presented a multi-agent based model of a Microgrid (formed by the interconnection of small, modular generation to low voltage distribution systems). Aside from selling energy to the network, the local distribution and generation units also have other tasks, like producing heat for local installations. These tasks showed the importance of a distributed control and autonomous operation. Dimeas et al. claimed that the use of multi-agent systems in the control of a microgrid solves a number of specific operational problems. Sweda et al. [4] created a simulation environment aimed at capturing the activities and decisions of an individual driver who has the option of purchasing an electric vehicle. In this agent based model, the agents are the drivers who can interact with each other to influence the vehicle purchasing behaviour. Each agent is given different attributes such as age, level of anxiety, income, etc. Vehicle maintenance costs have not been accounted for in this simulation; moreover agents know ahead of time when to replace their vehicles. This simulation helps investors in the PHEV industry acquire knowledge in the demand of electric vehicles in a specific area. Zhi et al. [5] tried to simulate an electricity market with PHEV penetration by using agents to simulate and model various situations. Each agent has a unique decision making process to plan their daily trips and determine vehicle profiles. The authors also tried to observe the vehicle owner's payments and investigate the relation to different market structures. They concluded that the current electric power structure is capable of accommodating the electric vehicles if appropriate charging strategies are applied. However, they do not account for a continuous increase of electrical vehicle agents within their simulation.

Vandael et al. [6] investigated fulfilling customer demands while avoiding infrastructure overloads by reducing imbalance costs. Hence, they proposed a Multi-Agent System (MAS) which utilizes an "intention graph" for expressing the flexibility of a fleet of PHEVs, and consequently help in scheduling their charging in real time to help reduce imbalances.

11.3 Technologies

The repast suite was used to implement the software agent paradigm. Multi-Agent Systems (MAS) was implemented in Repast in order to model the interactions between PHEVs, residential zones, industrial zones, grid stations and transformers. MAS are often used by researchers to achieve more individual intelligence and control [7]. We have also implemented "abstract" agents, which do not have a physical existence in the system but can manipulate and control other agents. One of such functionalities is increasing the number of PHEVs at specific time intervals, or assessing the loads on transformers, wires and grid stations. With respect to the simulated entities, the main agents behind the simulation structure are:

1. PHEVs
2. Residential and Industrial Zones
3. Grid Stations
4. Transformers
5. Abstract Agents.

11.4 Simulation Overview

We have used a total of 7 agents in our simulation. One of which is an abstract agent which does not have a physical existence in the system.

11.4.1 PHEVs

PHEVs are an integral part of any smart grid system, as such we have modelled the electric vehicle agents such that they are created every 100 ticks in the simulation world (equivalent to approximately one day in the real world). However, in addition to their tendency to move more often during the day than during the night (hence generating power usage that is close to their real-world counterparts) we have also created a more realistic model in which cars "jump" to office blocks (industrial agents in the simulation) from 8:00 a.m. to 5:00 p.m. In this manner, the

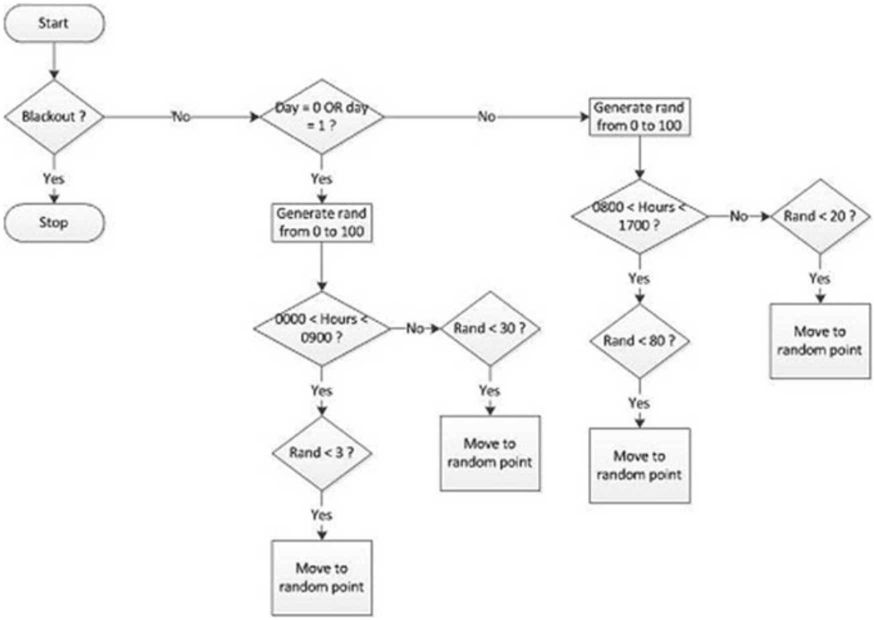


Fig. 11.1 Flow chart of PHEV movement algorithm

load transfers to the transformer assigned to the office block, which in turn transfers to the grid. The appropriate calculations to power consumption and production as well as transformer loads are calculated by the abstract agent. Figure 11.1 shows the movement of the electric cars.

11.4.2 Residential and Industrial Zones

In order to create a simulation as close to the real-world as possible, we have gone through multiple mappings of the city. Initially, we used a square function to distribute the residential and industrial cells across the map. As can be seen from Fig. 11.2, the residential cells are more concentrated towards the northeast while the industry cells are more concentrated towards the southwest. However, we soon discovered that this is not a common placement with big-city maps. Therefore for our next iteration we used a circular placement of agents for this implementation in which the outer rim of the circle is more populated by industrial agents, while the centre of the circle is more populated with residential agents. This is a more accurate representation of a real-world city since the more densely populated areas are usually inside the city, while the industrial zones, which contain factories that emit gasses into the air are usually on the outskirts, so as to preserve clean air for the residents (Fig. 11.3).

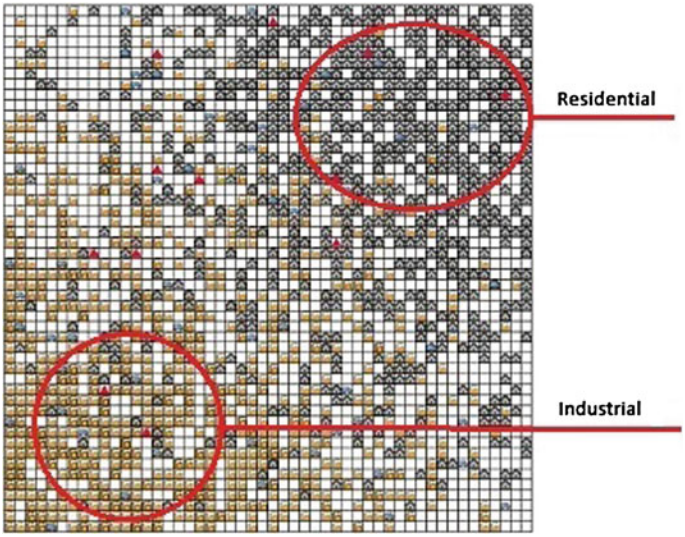


Fig. 11.2 Initial industrial and residential zone layout. *Note* how a chi-squared distribution function was used in order to distribute the residential agents and the industrial agents

Household agents are randomly generated blocks with randomly generated sizes. A transformer is placed on the left side of each block, providing only for that specific block. The transformer is directly connected to the grid, exchanging power with it back and forth according to demand, which in turn is regulated by the behaviour of PHEVs. Algorithm 1 illustrates the placement of households in the

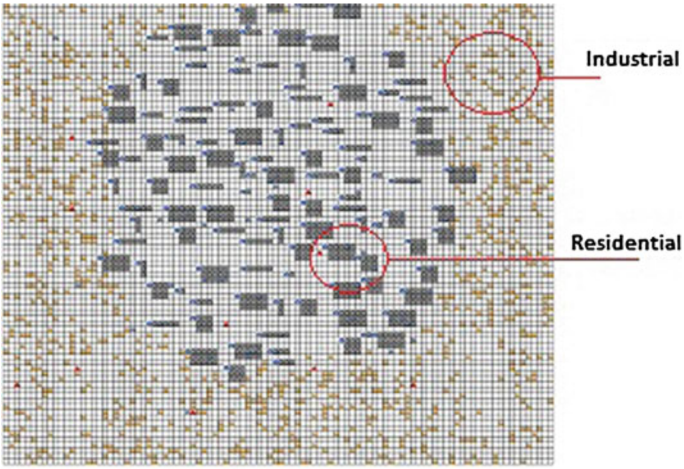


Fig. 11.3 New industrial and residential zone layout. *Note* how in addition to having the industrial districts on the outskirts of the city, the distribution of the city is now circular

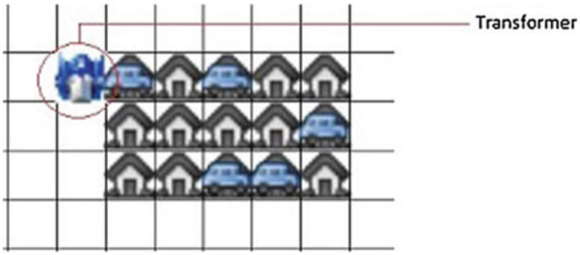


Fig. 11.4 New industrial and residential zone layout. *Note* how each block of residents now has a transformer unit attached

centre of the circle, as well as the placement of Transformers on the top left of each residential block (Fig. 11.4).

Calculations are first done to add each residential block (a block of house agents) in a circular fashion. The transformers are then added to the top left corner of the residential block, and the block is then considered its "child". This means that this transformer will be handling the power consumption and production of that residential block, in turn meaning it will be affected by the actions of the PHEVs belonging to that block. Algorithm 2 refers to the placement of industrial blocks (a block of industrial agents). It works in the same manner as Algorithm 1, but without the placement of transformers.

```
for all house agents do
    $occupancy \leftarrow 0$
    $blockWidth \leftarrow Ceil(Random(0,5))$
    $blockHeight \leftarrow Ceil(Random(0,2))$
    while $occupancy > 0$ do
        $occupancy \leftarrow 0$
        $\theta \leftarrow Random(0,360)$
        $x \leftarrow r \times (Cos((\theta \times (22/7))/180))\%(gridWidth + 1) + (gridWidth/2)$
        $y \leftarrow r \times (Sin((\theta \times (22/7))/180))\%(gridWidth + 1) + (gridWidth/2)$
        for all surrounding grid cells do
            $occupancy \leftarrow occupancy + residentialCellSize$
        end for
        $Add(TransformerAgent)$
    end while
    for $jj = -(blockWidth/2)$ to $(blockHeight/2)$ do
        for $ii = -(blockHeight/2)$ to $(blockHeight/2)$ do
            $Add(HouseAgent)$
            Place HouseAgent on the grid
            Add HouseAgent to transformer's Children (Transformer now becomes responsible
              for house)
        end for
    end for
end for
```

```
**for all** industrial agents **do**
    $occupancy \leftarrow 0$
    $blockWidth \leftarrow Ceil(Random(0,5))$
    $blockHeight \leftarrow Ceil(Random(0,2))$
    **while** $occupancy > 0$ **do**
        $occupancy \leftarrow 0$
        $\theta \leftarrow Random(0,360)$
        $x \leftarrow r \times (Cos((\theta \times (22/7))/180))\%(gridWidth + 1) + (gridWidth/2)$
        $y \leftarrow r \times (Sin((\theta \times (22/7))/180))\%(gridWidth + 1) + (gridWidth/2)$
        Place industrial agent at (x, y)
        **for all** surrounding grid cells **do**
            $occupancy \leftarrow occupancy + residentialCellSize$
        **end for**
    **end while**
**end for**
```

11.4.3 Grid Stations

For the purposes of this chapter, we have placed grid stations at random along the grid. This is to accommodate the possibility that the smart grid was not designed with optimal grid station placement in mind.

11.4.4 Transformers

A transformer is a device which transfers electrical energy from one circuit to another through inductively coupled conductors. We have set up our transformers such that we have one transformer per residential zone. The load from the residential zone passes from the residential agents to the transformer, which in turn passes it to the grid. In this way, the transformer only provides to the block that it is assigned to.

11.4.5 Abstract Agents

In addition to agents with a physical presence in the simulation, we have implemented two “abstract“ agents. These agents do not have a physical presence in the simulation, however they have access to information stored within the other agents, and thus are able to manipulate them. The first of these is an agent responsible for the electric vehicles in the simulation. Its sole responsibility is to add vehicles to the city at specified time intervals (the most practical interval for experiments was found to be every 100 ticks).

The second abstract agent is the stabilizer agent. This agent monitors the smart grid usage and changes things within the simulation accordingly. For example, if the grid is overloaded, the stabilizer will induce a blackout in an area. In addition to monitoring the other agents, this agent provides room for the expansion of the simulation in future work, as it allows for the monitoring of the entire system and the agents within it. Algorithm 3 describes the monitoring process and actions taken by the stabilizer when measuring the production and consumption. The stabilizer runs every step.

```
for each grid station $i_\alpha$ do
    $i_\alpha$.production = $i_\alpha$.consumption;
    if $i_\alpha$.production $>i_\alpha$.maxProductionLevel then
        $i_\alpha$.production = 0;
        $i_\alpha$.blackout = true;
        for each residential cell $j_\alpha$ do
            $j_\alpha$.blackout = true;
            for each vehicle associated with a house $k_\alpha$ do
                $k_\alpha$.blackout = true ;
            end
        end
        for each commercial cell $c_\alpha$ do
            $c_\alpha$.blackout = true;
        end
    end
end
```

11.5 Agent Mechanisms

The relationship of power consumption and production are different for the case of households and industries. In the case of households, we have set the grid and PHEVs as the both the destinations and the sources. For instance, power can be provided from the grid to the transformer, which in turn provides this power to the households, whose residents charge their PHEVs with this power. On the other hand when a user is charging their car, the load can pass from a PHEV to a household, in turn going to the transformer, which in turn transfers the load to the grid. The relation below shows the flow of power from the Grid to the PHEV and back.

$$Grid \leftrightarrow Transformer \leftrightarrow Houses \leftrightarrow PHEV \tag{11.1}$$

The case for industries is a simpler one in our simulation. The flow of power goes from the grid to the industrial zones directly. Once a PHEV arrives at an industrial agent, the power load transfers to the industrial agent. Otherwise, the load transfers to the house (Eq. 11.2).

$$Grid \leftrightarrow Industry \tag{11.2}$$

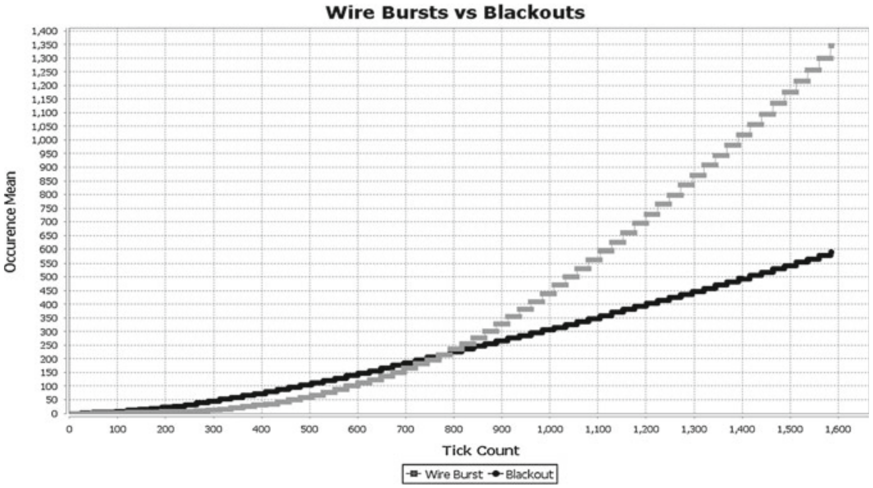


Fig. 11.5 Initial results. *Note* how the occurrence rate of the wire bursts exceeds that of the blackouts over time, while the blackouts continue to occur regularly in a linear fashion

11.6 Experiments and Results

We have used the city of Windsor, Ontario, Canada as a case study and experiment test bed. The Windsor area is approximately 146.91 km^2. Since our simulation contains 2,500 cells, we have calculated each cell to represent approximately 0.058764 km^2. According to EnWin utilities, Windsor has a power generation of 580 MW with an extra 8,580 MW, making the total power generation 9,160 MW. There is a total of 72 grid stations in the entire metro area of Windsor. However, by taking a ratio of the city area and scaling down the total size, we only needed to use 12 grid stations in our simulation, with each covering an area of 15 × 15 cells and producing approximately 763 MW.

In terms of residential and industrial zones, about 60 % (or approximately 1,500 cells) is industrial, while about 15 % are farms, and the remaining 35 % are houses. Hence, the residential area in Windsor is about 51.46 km^2, or 875 cells (we've included farms under residential zones). We have taken an average of 5 persons per family, with the density of Windsor being approximately $4,100/\text{km}^2$, our calculations show there are approximately 819 families/km^2. Out of this, 80 % of them would be in apartments, townhouses, etc. Hence our densities will be 163 houses/km^2 and 13 apartment buildings/km^2. Following these considerations, we have determined that there are 8 houses and 1 apartment building per cell in a household agent. Thus, we have 1,200 industrial cells, and 875 residential cells.

Finally, with regard to power consumptions, we have assumed that homes have a 50 kWh usage. Hence the approximate consumption by a residential cell is:

$$400\,(8\text{ houses})\ +1,5111\,\text{apartmentbuilding} = 1,911\ \text{KW} = 1.911\,\text{MWh} \tag{11.3}$$

While the power consumption by an average industrial cell can be taken as the difference between generation and consumption. Moreover, we have added an extra 1,000 MW for peak situations. Hence the average usage of an industrial cell is approximately 5.4 MW. Further, we have assumed that PHEVs consume about 30 KW for each charge. Therefore, we have assumed 1.911 mWh usage per residential cell, 5.4 MW usage per industrial cell, and 30 KW usage for each PHEV charge.

The simulation was run for 8,760 simulation-world hours, simulating approximately a year. Initial results (running the simulation without the use of transformer agents) show that there have been 10,817 wire bursts and 2,211 blackouts. Blackouts usually occur because of grid station failures while wire bursts typically occur because of wiring failures. Once a blackout or a wire burst occurs, we increment their corresponding variables and reflect this change on the simulation. We are thus able to extrapolate the next logical step in upgrading our smart grid. It is evident from our simulation that a wire upgrade is more important than a grid station upgrade to better utilize and sustain the current infrastructure for the city. Further, the simulation also presents when and where the blackouts and/or wire bursts will occur, further triangulating the problematic areas in order to allow for easier upgrade strategies.

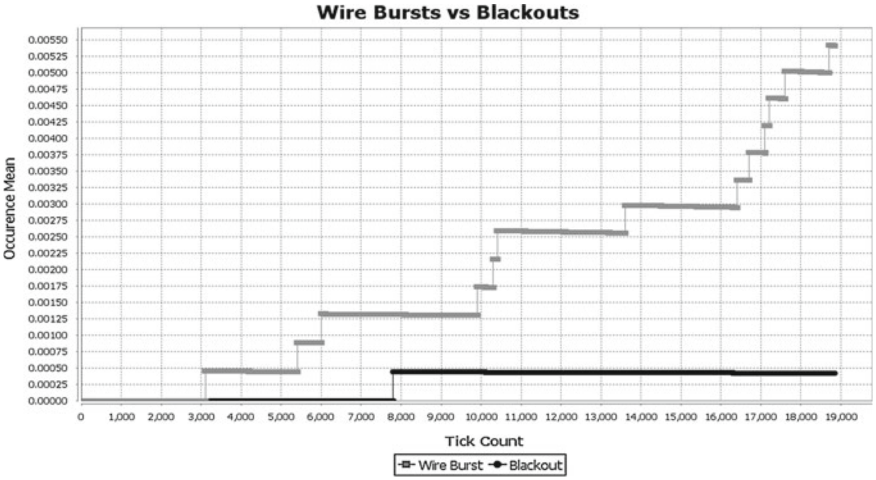


Fig. 11.6 New results. *Note* how the blackouts occur much less frequently now (in fact, only one occurrence at the time this snapshot was taken), and while wire bursts still occur more frequently, they still occur at a lesser rate than in our previous iteration

At the beginning, there were multiple blackouts in several residential regions, however the system stabilized after about 15 min (approximately 1 million ticks). We have found that there have been more wire bursts than there have been blackouts. In contrast to our initial results (Fig. 11.5), however, results using transformer agents show there have been no transformer failures, however blackouts and wire bursts are still occurring at relatively the same rate (Fig. 11.6).

The values may be changed and/or expanded to include more hardware upgrades and parameters. These results indicate that to realize a better equipped smart grid, a simulation is necessary, as changing the hardware (such as adding a transformer in this case) has changed our results drastically. This suggests that a numerical model to predict performance and usage may not be thorough or flexible enough.

11.7 Conclusion

We have presented a smart grid simulation model for a real-world city, and used a multi-agent system to better understand and make recommendations to a smart grid. However the previous model was lacking in both realism and hardware parameters. In this chapter we have extended this to a more realistic model which adopts the more realistic circular distribution of residential and industrial zones. By placing the residential zones in the inner circle, and the industrial zones in the outer circle, and assigning a transformer to each residential block we are able to achieve a more realistic simulation of a smart grid operated city.

Simulation results show that the multi-agent system approach is able to show when and where the blackouts, wire bursts or transformer failures will occur. They also show more stabilized results compared to our previous iteration [8], with the rate of occurrence of blackouts and wire bursts being much less linear than before. This indicates a more realistic result as the number of PHEVs increasing in the system should create many wire bursts, however once we added transformers the rate of black outs significantly stabilized. In the future, it is possible to expand the simulation to investigate how grid station placements will play a role on simulation results. It is also possible to include more hardware parameters and investigate how the results are influenced, possibly providing predictions as to which failure is likely to occur first and where.

Acknowledgments The authors acknowledge the contributions of the National Sciences and Engineering Research Council (NSERC).

References

1. Schneider KP, Tuffner FK (2010) Evaluation of conservation voltage reduction (CVR) on a national level. In: U.S. Department of Energy
2. James G, Cohen D, Dodier R, Platt G, Palmer D (2006) A deployed multi-agent framework for distributed energy applications. In: Proceedings of the fifth international joint conference on autonomous agents and multiagent systems. AAMAS '06. ACM. 1-59593-303-4. Hakodate, Japan
3. Dimeas A, Hatziargyriou N (2004) A multiagent system for microgrids. In: IEEE power engineering society general meeting, vol 1, pp 55–58
4. Sweda T, Klabjan D (2011) An agent-based decision support system for electric vehicle charging infrastructure deployment. In: IEEE vehicle power and propulsion conference (VPPC)
5. Zhi Z, Jianhui W, Botterud A (2011) Agent-based electricity market simulation with plug-in hybrid electric vehicle penetration. In: IEEE power and energy society general meeting, July 2011. 1944-9925, pp 1–2
6. Vandael S, Boucké N, Holvoet T, De Craemer K, Deconinck G (2011) Decentralized coordination of plug-in hybrid vehicles for imbalance reduction in a smart grid. In: The 10th international conference on autonomous agents and multiagent systems, vol 2. AAMAS '11. 0–9826571-6-1, 978–0-9826571-6-4. Taipei, Taiwan
7. Karnouskos S, de Holanda TN (2009) Simulation of a smart grid city with software agents. In: Third UKSim European symposium on computer modeling and simulation. EMS '09. pp 424–429
8. Shaabana A, Syed S, Kobti Z, Tepe K (2012) Decision support using a multi-agent system for hardware upgrades in smart grids. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, USA, San Francisco, 24–26 Oct 2012, pp 1287–1292

Chapter 12
Ballistic Behaviour in Bounded Velocity Transport

F. Debbasch, D. Espaze and V. Foulonneau

Abstract Stochastic models of bounded velocity transport are revisited. It is proven that these models exhibit short-time propagative (as opposed to diffusive) behavior for a large class of initial conditions. Numerical simulations also show that this propagative effect is different from the damped propagation predicted by common hyperbolic models. A fit of the density profiles is finally presented and a geometrical generalization of Fick's law is also proposed.

Keywords Diffusions · Fick's law · Geometric flows · Hyperbolic diffusion · Relativistic Ornstein-Uhlenbeck process · Stochastic processes

12.1 Introduction

Transport at bounded velocity is encountered in many different contexts which range from metal [1, 2] and computer engineering [3] to tumor treatment [4, 5] and fusion plasma physics [6, 7]. Finding realistic models of such transport is a long standing problem of continuous media theories [8–12]. The simplest form of transport is matter transport and the simplest models of matter transport are stochastic processes. Stochastic processes which bound the velocity of the diffusing matter have been introduced in [9]. We show here by an analytical computation that, for a very large and natural class of initial conditions, these processes display propagative (ballistic)

F. Debbasch (✉) · D. Espaze · V. Foulonneau
LERMA, UMR 8112, UPMC, Paris, France
e-mail: fabrice.debbasch@gmail.com

D. Espaze
e-mail: davidespaze@yahoo.fr

V. Foulonneau
e-mail: vfoulonneau@hotmail.fr

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_12,

behaviour at short times [13]. This propagation (ballistic) effect contrasts sharply with (1) standard diffusive behaviour, which only appears at asymptotic large times (2) damped propagation predicted by the widely used hyperbolic transport models based on the telegraph equation. We illustrate these findings by numerical simulations of the ROUP [14], which is the first bounded velocity process introduced in the physical literature. We also present a simple analytical *Ansatz*, which fits the density of the ROUP at all times to an accuracy of order 3 % and which can be used in a more general context to model bounded velocity transport. We finally propose a generalization of Fick's law to bounded velocity diffusions. This generalization is based on a simple geometrical models and connects bounded velocity transport to other fields of mathematics and physics, and in particular to the theory of geometrical flows.

12.2 Short-Time Propagative Behaviour of Bounded Velocity Processes

Consider the following 1D stochastic process:

$$dx_t = v_t \, dt \tag{12.1}$$

$$dv_t = F(v_t)\, dt + \boldsymbol{\sigma}(v_t)\, dB_t \tag{12.2}$$

where F is a friction or dissipative term and $\boldsymbol{\sigma}$ is a noise coefficient. Equation (12.1) is simply the definition of the velocity v as the time-derivative of the position x and (12.2) is a generalization of Langevin equation.

Since we are modeling bounded velocity transport, we suppose that the initial condition and the process itself restricts v to a finite interval, say $I = (-c, +c)$, where c is an arbitrary constant velocity which depends on the nature of the diffusing particles and of the medium in which transport occurs. A simple one-to-one map of this interval onto $\mathbb{R}$ is of course:

$$\begin{aligned} p : I &\rightarrow \mathbb{R} \\ v &\rightarrow p(v) = \gamma(v)\, v \end{aligned} \tag{12.3}$$

with $\gamma(v) = (1 - \frac{v^2}{c^2})^{-1/2}$. Note that $v \rightarrow \pm c$ corresponds to $p \rightarrow \pm\infty$ and $\gamma \rightarrow +\infty$. The first equation of the process transcribes into

$$dx_t = \frac{p_t}{\Gamma(p_t)}\, dt \tag{12.4}$$

with $\Gamma(p) = (1 + p^2/c^2)^{1/2}$. The variable p will henceforth be called the momentum of the diffusing particle.

Consider a diffusing particle starting its motion from point $x_0 = 0$ with initial momentum p_0. For sufficiently small times, the position varies with time according to:

$$x_t = \frac{p_0}{\Gamma(p_0)} t + O(t^2). \tag{12.5}$$

The probability law of x_t and, thus, the density n are then entirely determined by the probability law of p_0 i.e. the initial momentum distribution of the particle, which we denote by $F^*(p)dp$. To make the discussion simpler, suppose that F^* isotropic, and write $F^*(p) = \exp(-\Phi(\Gamma(p)))$.

Equations (12.4) and (12.5) show that x_t/t has the same probability law as the initial velocity of the particle. This law can be obtained by changing variables in the initial momentum distribution $F^*(p)$. By direct differentiation,

$$dp = (\gamma(v))^3 dv \tag{12.6}$$

and this leads to

$$n(t,x) \approx \frac{1}{t}\left(\gamma\left(\frac{x}{t}\right)\right)^3 \exp\left(-\Phi\left(\gamma\left(\frac{x}{t}\right)\right)\right). \tag{12.7}$$

The maxima and minima of n at short time can be identified by computing the first and second derivatives of this expression with respect to x. One finds that, for all increasing function Φ such that the equation $\gamma\Phi'(\gamma) = 3$ admits a single solution γ^*, the points $\pm x^*(t) = \pm c^* t$ with $c^* = c\sqrt{1 - 1/\gamma^{*2}}$ are maxima of the density provided the function Φ is convex.[1] Thus, for any convex function Φ, the short-time density profile exhibits two peaks which travel at constant velocity c^*. The short-time transport thus exhibits propagative (ballistic) behaviour.

12.3 Illustration Through Numerical Simulations

The ROUP [14, 15] corresponds to the choices $F(v) = -\alpha v$ and $\boldsymbol{\sigma}(v) = \sqrt{2D}$, where α and D are both constant. The constant α is a friction coefficient, and D is a noise coefficient. Numerical simulations have been performed for the one-parameter family of functions Φ:

$$\Phi_\beta(\gamma) = \beta\gamma + a(\beta), \tag{12.8}$$

where β is an arbitrary real positive coefficient and $a(\beta)$ ensures that the corresponding initial momentum distribution F^*_β is normalized to unity with respect to dp. This distribution is a hyperbolic distribution commonly called Jüttner distribution [16, 17] and β plays the role of an initial inverse temperature for the ROUP. At fixed p and thus, at fixed value of $\Gamma(p)$, the ratio $A_\beta(p) = F^*_\beta(p)/F^*_\beta(0)$ decreases exponentially as β increases. For $\beta \gg 1$, $A_\beta(p)$ is comparable to unity

[1] This last condition is sufficient but not necessary for the extrema at $\pm x^*(t)$ to be maxima. The necessary and sufficient condition is $D^* = -3\gamma^* - \gamma^{*3}\Phi''(\gamma^*) < 0$.

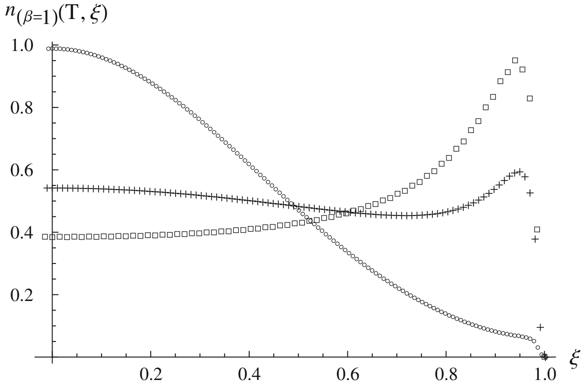


Fig. 12.1 Density profile and propagation. Density-profile n_β against rescaled position $\xi = x/(ct)$ for $\beta = 1$ and $T = \alpha t = 0.5$ (*squares*), $T = 2$ (*crosses*) and $T = 10$ (*circles*). The density n_β is normalized to unity with respect to $d\xi$

only if $p \ll c$. This means that the diffusing particle, initially, does not 'see' the maximum velocity c; note that, for $p \ll c$, the function $\Gamma(p)$ can be approximated by its expansion around $p = 0$ i.e. $\Gamma(p) \simeq 1 + p^2/(2c^2)$ and F^*_β is then approximately Gaussian.

Figures 12.1 and 12.2 display typical profiles for the density n_β generated by the ROUP with initial condition (12.8). Figure 12.1 displays $n_{\beta=1}$ as a function of $\xi = ct$ for different values of the dimensionless time $T = \alpha t$.

At early times, the maximum of the density profile is not situated at $\xi = 0$ i.e. at the starting point of the diffusion, but rather at $\mid \xi_{\beta=1} \mid \approx 0.948$, remarkably close to the analytical prediction $\mid x^*_\beta(t)/(ct) \mid = \mid \xi^*_\beta(t) \mid = 0.943$ (see Sect. 12.2). In time, a secondary maximum appears at the origin point $\xi = 0$. This secondary maximum grows and finally becomes much higher than the peaks at $\pm\xi_{\beta=1}$. The density profile thus gets closer and closer to a Gaussian and the bounded velocity transport transforms into standard diffusion with diffusion coefficient $\chi = D/\alpha^2$ in physical space, as expected from [18, 19].

Fix now an arbitrary, not necessarily large dimensionless time T and consider the density profile $n_\beta(T, \cdot)$ at this time. As β tends to infinity, this density profile

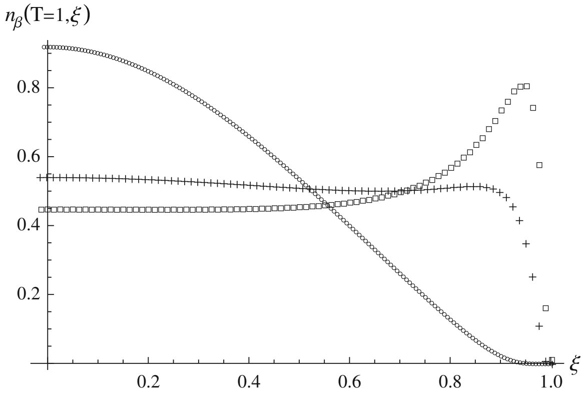


Fig. 12.2 Density profile and propagation. Density-profile n_β against rescaled position $\xi = x/(ct)$ for $T = 1$ and $\beta = 1$ (*squares*), $\beta = 1.2$ (*crosses*) and $\beta = 2$ (*circles*). The density n_β is normalized to unity with respect to $d\xi$

also tends towards the standard diffusion Gaussian. Indeed, the more β increases, the less the diffusing particle 'sees' the velocity bound c (see the discussion above) and the more the transport looks like standard Fickian diffusion. Note also that the time at which the secondary maximum appears at the origin is a decreasing function of β and tends to zero as β tends to infinity.

12.4 Failure of the Hyperbolic Diffusion Model

Let us show that the spatial density of the ROUP does not obey Cattaneo's hyperbolic diffusion equation [8], which is a popular model of bounded speed transport. Cattaneo's damped wave equation reads:

$$\partial_t n = \chi\left(\partial_{xx} - \frac{1}{c^2}\partial_{tt}\right)n, \tag{12.9}$$

where $\chi = D/\alpha^2$ is the usual diffusion coefficient in position space. We have computed numerically the relative discrepancy $R_\beta(T)$ between $\partial_t n_\beta$ and $\chi\left(\partial_{xx} - \frac{1}{c^2}\partial_{tt}\right)n$; Fig. 12.3 displays a typical result in Fourier space. This figure clearly displays the failure of Cattaneo's hyperbolic diffusion model to reproduce the correct density profile of the ROUP.

12.5 Fit of the Density Profile

Consider now the following function of position and time:

$$N_{\alpha,\sigma,B}(t,x) = B(t)\left(\gamma\left(\frac{x}{t}\right)\right)^{\alpha(t)} \exp\left(-\phi_Q\left(\gamma\left(\frac{x}{t}\right)\right)\right) \times \exp\left(-\frac{x^2}{2\sigma(t)^2}\right), \tag{12.10}$$

where α and σ are two arbitrary functions of t and B ensures that $N_{\alpha,\sigma,B}$ is at all times normalized to unity on $(-ct, ct)$.

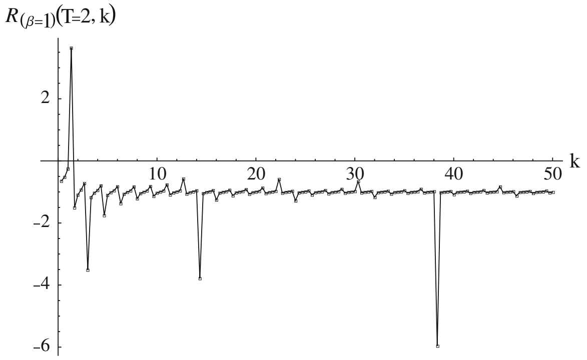


Fig. 12.3 Failure of the hyperbolic diffusion model. Evolution with k of the relative discrepancy between $\partial_t n_\beta$ and $\chi\left(\partial_{xx} - \frac{1}{c^2}\partial_{tt}\right)n_\beta$ for $\beta = 1$ and $T = 2$. The hyperbolic Cattaneo model predicts R identically vanishes

Fig. 12.4 Fit of the density profile. Time-evolution of the α-coefficient for $\beta = 1$

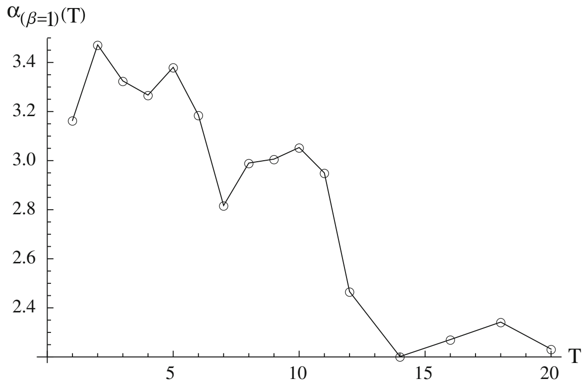
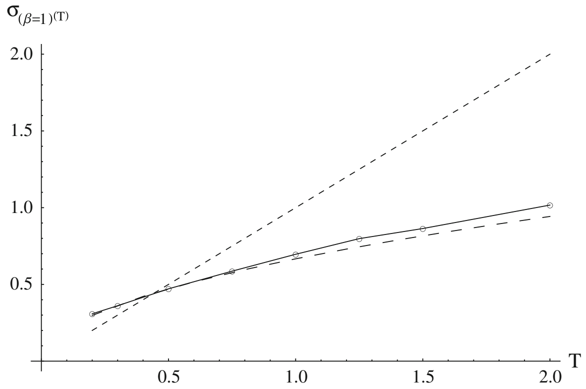


Fig. 12.5 Fit of the density profile. Time-evolution of the σ-coefficient (*circles*) for $\beta = 1$. The *straight line* is $x = ct$ and the *dashed curve* is $x = (2/3)\chi\sqrt{t}$



At each time t, the values of $\alpha_\beta(t)$, $\sigma_\beta(t)$ and $B_\beta(t)$ producing the best fits of n_β can be obtained, for example, by minimizing the following distance function $d_C(t)$ between $n_\beta(t,\cdot)$ and $N_{\alpha\sigma B}(t,\cdot)$:

$$\begin{aligned} d_C(T) = &\int_{\mathbb{R}} \mid n_\beta(t,x) - N_{\alpha\sigma B}(t,x) \mid dx \\ &+ \lambda_B \mid 1 - \int_{\mathbb{R}} N_{\alpha\sigma B}(t,x) dx \mid, \end{aligned} \tag{12.11}$$

where λ_B is a Lagrange multiplier which enforces normalization and the convention $N_{\alpha\sigma B}(t,x) = 0$ for $\mid x \mid\ > ct$ has been used.

Figures 12.4, 12.5 and 12.6 display the results of the fit at different times for $\beta = 1$ and $\lambda_B = 100$. The precision of the fit is always better than $3\,\%$ (see Fig. 12.6). The coefficient α_β remains close to the above computed value of 3 and seems to globally decrease with time; its average value for the points plotted in Fig. 12.4 is 2.88. At small times, $\sigma_\beta(t)$ behaves like $2\sqrt{t}/3$ (see the green dashed curve in Fig. 12.5) and is thus larger than ct (the red straight line in Fig. 12.5); the Gaussian then varies slowly over $(-ct, +ct)$, the shape of the density profile is

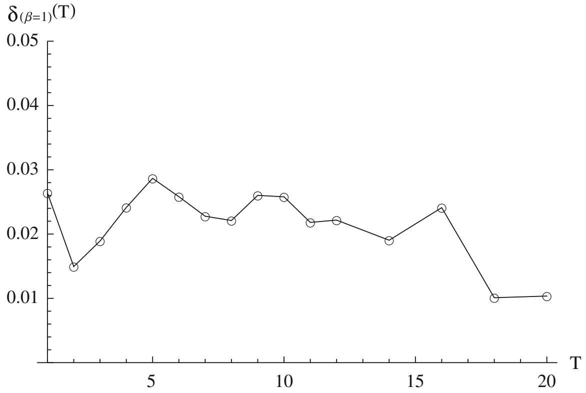


Fig. 12.6 Fit of the density profile. Time-evolution of the absolute error of the fit for $\beta = 1$

essentially controlled by the $\gamma^{\alpha_\beta(t)} \exp(-\beta\gamma)$ and it thus displays the characteristic peaks x close to ct. As t increases, $\sigma_\beta(t)$ becomes smaller than ct and the maximum of the Gaussian at $x = 0$ generates the secondary maximum at $x = 0$. As time still increases, $\sigma_\beta(t)$ increases slowly from $2\chi\sqrt{t}/3$ to $\chi\sqrt{t}$ (the onset of this increase can actually be seen in Fig. 12.5) but $\sigma_\beta(t)/(ct)$ continues to decrease towards zero; the density profile is then essentially controlled by the Gaussian and tends towards the standard result predicted by Fick's law.

Let us stress that the fit presented in this section is not based on an approximate analytical computation of the finite-time density profile, but is only a heuristic extension of the short-time computation presented in the previous section. This fit nevertheless highlights the fact that the whole time-evolution of the density profile can be understood in very simple terms i.e. as the superposition of two competing phenomena which are (1) the propagation of the peaks at velocity close to the light-velocity (2) a standard Gaussian diffusion with a typical scaling as $\sqrt{t}$.

The fit also constitutes a simple, ready-to-use model of finite speed transport. It can be easily integrated into numerical simulations and should thus prove useful in a wide variety of physical and engineering applications.

12.6 Fick's Law for Bounded Velocity Diffusions

Standard Fick's law cannot be used to model the density profiles presented in Sect. 12.3. Indeed, standard Fick's law predicts that the density profile of a particle starting its diffusive motion from a given point in space will be Gaussian at all times, and thus does not reproduce the ballistic effect observed on Figs. 12.1 and 12.2.

We will now present a generalization of Fick's law for bounded velocity diffusions. This generalization is based on a geometrical model first introduced in [20, 21] to describe diffusions on interfaces with time-varying geometries. Let us therefore start by recollecting a few facts about diffusions on general surfaces or membranes.

Locally, points on a surface can be charted by two coordinates, say $x = (x^1, x^2)$. The geometry of the surface is entirely described by the so-called metric g, which defines the distance between two infinitesimally adjacent points. Given a choice of coordinates, the metric is represented by its components $g_{ij}(t,x)$, $(i,j) \in \{1,2\}^2$, and the line-element dl around point x at time t is $dl_g^2 = g_{ij}(t,x)dx^i dx^j$, where summation over i and j is understood (Einstein summation convention).

The metric also defines an infinitesimal area element dA on the surface:

$$dA_g = \boldsymbol{\sigma}(t,x)dx^1 dx^2, \tag{12.12}$$

where $\boldsymbol{\sigma}(t,x) = \sqrt{\det\, g_{ij}(t,x)}$ is the determinant of the metric components g_{ij}. The standard Lebesgue measure $dx^1 dx^2$ is actually the area-element dA_η generated by the Euclidean, fat and time-independent metric $\boldsymbol{\eta}$ with components $(\eta_{ij}) = \text{ diag } (1,1)$.

Consider now a diffusion on the surface. A natural object is the density n^g of this diffusion with respect to the area element dA_g. The evolution equation obeyed the density n^g of a Brownian motion defined on an arbitrary surface has been introduced in [22, 23] for surfaces with constant geometries (see also [24] for a pedagogical account) and has been generalized in [20, 21, 25] to situations where the geometry of the surface varies with time. The general diffusion equation for Brownian motion of a surface with time-dependent geometry reads:

$$\partial_t(\boldsymbol{\sigma}(t,x)n^g(t,x)) = \partial_i\left(g^{ij}(t,x)\boldsymbol{\sigma}(t,x)\partial_j n^g(t,x)\right), \tag{12.13}$$

where ∂_i stands for the derivative with respect to x^i and g^{ij} is the matrix inverse to g_{ij}. This equation naturally conserves the 'total particle number' $\mathcal{N} = \int_x n^g(t,x)dA_g$.

The same Brownian motion can also be described by its density $n^\eta(t,x)$ with respect to dA_η. Since $dA_g = \boldsymbol{\sigma}(t,x)dA_\eta$, $n^\eta(t,x) = \boldsymbol{\sigma}(t,x)n^g(t,x)$ and the diffusion equation for n_η reads:

$$\partial_t n^\eta(t,x) = \partial_i\left(g^{ij}(t,x)\boldsymbol{\sigma}(t,x)\partial_j\left(\frac{n^\eta(t,x)}{\boldsymbol{\sigma}(t,x)}\right)\right), \tag{12.14}$$

The idea behind the generalization of Fick's law to diffusions with bounded velocity is to consider these diffusions as transport in a non-trivial time-dependent metric. Let us develop this idea for the $1D$ diffusions considered in this article. Switching to the notations used in the previous sections, Eq. (12.15) transcribes into $\partial_t n + \partial_x j = 0$ with $n = n^\eta$ and

$$j(t,x) = \frac{1}{\boldsymbol{\sigma}(t,x)}\,\partial_x\left(\frac{n(t,x)}{\boldsymbol{\sigma}(t,x)}\right), \tag{12.15}$$

where x is now the standard coordinate on the real line.

Consider now the density profiles n_β presented in Sect. 12.3 and the associated currents j_β. For each β i.e. for each couple (n_β, j_β), Eq. (12.16) can be viewed as a

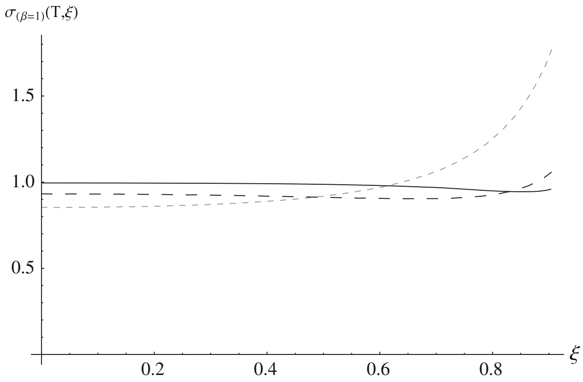


Fig. 12.7 Diffusion function σ_β appearing in the generalized Fick's law. Function σ_β plotted against the rescaled variable $\xi = x/(ct)$ for $\beta = 1$ and $T = 1$ (*continuous curve*), $T = 4$ (*small dashes curve*), $T = 10$ (*large dashes curve*). Note that traditional Fick's law is recovered as β tends to infinity and σ_β approaches a constant value

differential equation to be solved for an effective diffusion function σ_β. An explicit solution of (12.16) is:

$$\frac{1}{\boldsymbol{\sigma}_\beta^2(t,x)} = -2\int_{-ct}^{x} n_\beta(t,y)j_\beta(t,y)dy. \tag{12.16}$$

The function $\boldsymbol{\sigma}_\beta$ defines the effective difusion metric $g_\beta = \boldsymbol{\sigma}_\beta^2$ and, thus, the effective line-element $dl_\beta^2 = \boldsymbol{\sigma}_\beta^2(t,x)dx^2$ at time t around point x.

The standard, Galilean Relativistic Ornstein-Uhlenbeck Process corresponds to $c = +\infty$; as shown in [26], $j_\infty = -\chi(t)\partial_x n_\infty$, so that $\boldsymbol{\sigma}_\infty(t,x) = 1/(\boldsymbol{\chi}(t))^2$, which does not depend on the position x. The function $\boldsymbol{\chi}(t)$ also tends to a constant as t tends to infinity [26] and one thus recovers the standard Fick's law as both c and t tend to infinity.

Typical results are displayed in Fig. 12.4. The function $\boldsymbol{\sigma}_\beta$ is nearly flat at the centre of the interval $(-ct, ct)$ but grows to infinity near $\mid x \mid\ \sim ct$. Consider for example a point with coordinate x and a point with coordinate $x + \Delta x$, $\mid \Delta x \mid \ll \mid x \mid$. As far as the diffusion is concerned, the effective distance between these two points at time t is $\sigma_\beta(t,x)\Delta x$. This distance grows to infinity for any finite Δx when x approaches $\pm ct$. Thus, the distance that a particle needs to travel to get closer to ct by the amount Δx tends to infinity as the particle approaches $\pm ct$. This prevents the particle from ever crossing $c = \pm ct$ i.e. from being transported at velocities higher than c.

12.7 Discussion

We have proved that bounded velocity diffusions exhibit short-time propagative behaviour for a wide class of initial conditions. This has been illustrated by numerical simulations of the ROUP. We have also shown numerically that the widely used hyperbolic diffusion model does not replicate the density profiles of the ROUP and we have also presented a simple *Ansatz* which fits these profiles to a precision better than 3 %. We have finally proposed a geometrical generalization of Fick's law to bounded velocity diffusions.

We believe several general conclusions can be drawn from these results. First, the failure of the hyperbolic diffusion model to replicate the density profile of the ROUP, which is certainly the simplest reliable model of bounded velocity transport, strongly suggests that this model also fails in more complicated problems involving mass transport as well as momentum viscous transport and heat conduction.

The material presented in this article indicates (1) that bounded velocity effects are essentially short-time effects (2) that these effects depend on all characteristics of the initial state of the system in which transport is to occur. Indeed, in the situation studied in this article, the short-time density profile depends, not only on the initial position of the diffusing particle, but also on its initial velocity distribution and, in particular, on its initial temperature.

Let us also metion that the geometrical generalization fo Fick's law points to connections between bounded velocity transport and various very active fields of mathematics and physics. These include [26] geometrical flows and black hole thermodynamics.

Finally, all the results presented in this article need to be extended properly to include viscous momentum transfer and heat conduction. This can be accomplished, at least in theory, by analyzing kinetic models richer than the ROUP which also bound particle velocities.

Acknowledgments Part of this work was funded by the ANR Grant 09-BLAN-0364-01.

References

1. Itina TE, Mamatkulov M, Sentis M (2005) Nonlinear fluence dependencies in femtosecond laser ablation of metals and dielectrics materials. Opt Eng 44(5):051109–051116
2. Klossika JJ, Gratzke U, Vicanek M, Simon G (1996) Importance of a finite speed of heat propagation in metal irradiated by femtosecond laser pulses. Phys Rev B 54(15):10277–10279
3. Chen HT, Song JP, Liu KC (2004) Study of hyperbolic heat conduction problem in IC Chip. Japanese, J Appl Phys 43(7A):4404–4410
4. Jaunich MK et al (2006) Bio-heat transfer analysis during short pulse laser irradiation of tissues. Intl J Heat Mass Transf 51:5511–5521
5. Kim K, Guo Z (2007) Multi-time-scale heat transfer modeling of turbid tissues exposed to short-pulse irradiations. Comput Methods Programs Biomed 86(2):112–123
6. Freidberg J (2007) Plasma physics and fusion energy. Cambridge
7. Martin-Solis JR et al (2006) Enhanced production of runaway electrons during a disruptive termination of discharges heated with lower hybrid power in the frascati tokamak upgrade. Phys Rev Lett 97:165002
8. Cattaneo C (1948) Sulla conduzione del calore. Atti Sem Mat Fis Univ Modena, 3.
9. Chevalier C, Debbasch F, Rivet JP (2008) A review of finite speed transport models. In: Proceedings of the second international forum on heat transfer (IFHT08), 17–19 Sept (2008), Tokyo, Japan, Heat Transfer Society of Japan

10. Herrera L, Pavon D (2001) Why hyperbolic theories of dissipation cannot be ignored: comments on a paper by Kostadt and Liu. Phys Rev D 64:088503
11. Israel W (1987) Covariant fluid mechanics and thermodynamics: an introduction. In: Anile A, Choquet-Bruhat Y (eds) Relativistic fluid dynamics. Lecture notes in mathematics, vol 1385. Springer, Berlin.
12. Müller I, Ruggeri T (1993) Extended thermodynamics. Springer Tracts in Natural Philosophy, vol 37. Springer, New-York
13. Debbasch F, Espaze D, Foulonneau V (2012) Novel aspects of bounded veloity transport. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS 2012, USA, San Francisco, 24–26 Oct 2012, pp 1198–1201
14. Debbasch F, Mallick K, Rivet JP (1997) Relativistic Ornstein-Uhlenbeck process. J Stat Phys 88:945
15. Debbasch F, Chevalier C (2006) Relativistic stochastic processes. In: Proceedings of XV conference on non-equilibrium statistical mechanics and nonlinear physics, Mar del Plata, Argentina, 4–8 Dec 2006, A.I.P. Conference Proceedings, 2007
16. Debbasch F (2008) Equilibrium distribution function of a relativistic dilute perfect gas. Process Phys A 387:2443–2454
17. Jüttner F (1911) Das Maxwellsche Gesetz der Geschwindigkeitsverteilung in der Relativtheorie. Ann Phys (Leipzig) 34:856
18. Angst J, Franchi J (2007) Central limit theorem for a class of relativistic diffusions. J Math Phys 48(8)
19. Debbasch F, Rivet JP (1998) A diffusion equation from the relativistic Ornstein-Uhlenbeck process. J Stat Phys 90:1179
20. Chevalier C, Debbasch F (2007) Multi-scale diffusion on an interface. Eur Phys Lett 77:20005–20009
21. Chevalier C, Debbasch F (2008) Is brownian motion sensitive to geometry fluctuations? J Stat Phys 131:717–731
22. Itô K (1950) On stochastic differential equations on a differentiable manifold. Nagoy Math J 1:35–47
23. Itô K (1953) On stochastic differential equations on a differentiable manifold ii. M.K. 28:82–85
24. Øksendal B (1998) Stochastic differential equations, 5th edn. Universitext, Springer, Berlin
25. Chevalier C, Debbasch F (2010) Lateral diffusions: the influence of geometry fluctuations. Eur Phys Lett 89(3):38001
26. Debbasch F, Di Molfetta G, Espaze D, Foulonneau V (2012) Propagation in quantum walks and relativistic diffusions. Accepted for publication in Physica Scripta

Chapter 13
Fuzzy Logic Control Versus Traditional PI Control Applied to a Fixed Speed Horizontal Axis Wind Turbine

Luis Alberto Torres Salomao, Hugo Gámez Cuatzin, Juan Anzurez Marín and Isidro Ignacio Lázaro Castillo

Abstract A comparison between three types of control algorithms for a 1.5 MW horizontal axis fixed speed wind turbine is presented. A fuzzy logic proportional integral control (Fuzzy PI), a fuzzy logic control (FLC) and a classical proportional integral (PI) control are tested. A robustness test by adding noise to the wind speed signal is also performed. Design of the proposed Fuzzy PI control algorithm was achieved via tuning with the Ziegler-Nichols approach, using the same methodology for the PI controller tuning with the difference of incorporating a fuzzy logic section. The fuzzy logic section selects the desired PI gains according to wind speed with a smooth control transition. Fuzzy logic control was designed to obtain maximum power extraction at low wind speeds and to limit power extraction at a 1.5 MW nominal power set point. Aerodynamic characteristics of the wind turbine were studied in order to gain a basic understanding of the system dynamics. A 1.5 MW horizontal axis wind turbine model was designed for tuning as well as simulation performance studies. Results demonstrate the effectiveness of all techniques, achieving a controlled power extraction near the nominal value for the three controllers and maximum power extraction in low wind speeds for the Fuzzy PI and FLC control algorithms.

L. A. Torres Salomao (✉)
ACSE, University of Sheffield, D09 Amy Johnson Building, Portobello Street,
Sheffield S1 3JD, UK
e-mail: latsalomao@ieee.org

H. Gámez Cuatzin
CIDESI, CONACyT, 702 Av. Playa Pié de la Cuesta, 76130 Desarrollo San Pablo,
Santiago de Querétaro, Mexico
e-mail: hgamez@cidesi.mx

J. Anzurez Marín · I. I. Lázaro Castillo
FIE, Universidad Michoacana de San Nicolás de Hidalgo, Edificio Ohmega 2 Ciudad
Universitaria, Av. Fco. J. Mújica, 58040 Felícitas del Río, Morelia, Mexico
e-mail: j.anzurez@ieee.org

I. I. Lázaro Castillo
e-mail: i.lazaro@ieee-sco.org

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_13,

Keywords Fuzzy logic control · Non-linear control · Proportional integral control · Renewable energy production · Wind power · Wind turbine

13.1 Introduction

Control algorithm design for renewable energy production is a topic of great concern nowadays because of all the efforts around the mitigation of greenhouse effects. Many countries have modified their energy production plans for the near future by advancing green energy technologies via government funding and tax reductions [1]. Wind energy is the technology with the most rapid growth, but since wind is an intermittent resource, efficiency of this machine is of outmost importance.

In Mexico, almost all installed wind turbines are of foreign design and construction, being fixed speed horizontal axis wind turbines the most common. Mexico presents zones like in *La Ventosa*, *Oaxaca* with wind persistence and speeds that make them some of the best installation sites in the world.

In practice, the majority of the installed wind turbines have pitch control systems with traditional Proportional-Integral (PI) algorithms. These control systems are designed near the nominal wind speeds and power extraction values because of their good response for linear model systems as well as their implementation simplicity. However, the dynamic properties of large wind turbines make them highly non-linear systems, and in order to obtain maximum power extraction, non-linear control algorithms are required.

Fuzzy Logic Control (FLC) technique has been used for over 20 years with many successful applications like in [2–5]. It is an ideal technique for complex systems that are difficult to model or that present important parameter variation [4]. FLC design focuses in gaining a basic understanding of the plant in order to design an appropriate set of rules that can be directly loaded into the fuzzy controller. This is completely opposite to a traditional PI control, where focus is on modeling and the use of this model [6].

However, FLC applied for wind turbine applications, where fine control action is needed shows no robustness characteristics when dealing with important wind speed input noise. PI control schemes usually concentrate in reducing the error generated from the desired nominal power extraction value minus the actual power extraction and work well near nominal wind speeds. Their downfall is when analyzed for low wind speeds where non-linear action is needed to achieve maximum power extraction.

A Fuzzy PI control algorithm that combines the fuzzy logic control direct non-linear characteristics as well as the PI control effectiveness for power extraction error reduction is presented [3, 5]. Performance curves were analyzed to obtain a basic understanding of the wind turbine as well as maximum power capabilities at low wind speeds. An understanding of the optimum way to control the wind turbine was obtained from a direct analysis of these performance curves.

13.2 Wind Turbine Model

A wind turbine model is basically constructed with a mechanical turbine (*low speed rotor and blades*), gearbox (*multiplicative*) and the electric generator (*high speed rotor*) as can be appreciated at Fig. 13.1.

A wind turbine is a device designed to extract kinetic energy from wind [7]. When designing a wind turbine it is important to define the amount of energy to be extracted [7]. Available wind power is the time derivative of this kinetic energy:

$$P_v = \frac{dK}{dt} = \frac{1}{2}\rho A v^2 \frac{dx}{dt} = \frac{1}{2}\rho A v^3 \tag{13.1}$$

where P_v (W) is wind power, K (J) is kinetic energy available at the wind, ρ (kg/m^3) is air density, v (m/s) is wind speed, A (m^2) is the wind parcel's sectional area and x (m) its width.

This equation represents the amount of energy theoretically available for extraction. However, a limit exists in the extractable energy. This limit is defined as the power coefficient C_p dependent on the wind turbine aerodynamics. The maximum C_p available for extraction is known as the Betz limit, and to date, no wind turbine has been able to exceed it. Maximum achievable C_p according to Betz limit is $C_{p\ lim} = 0.593$ [7].

13.2.1 Mechanical Turbine

The mechanical turbine is the aerodynamically designed element to extract power from the wind and to communicate this power to the multiplicative gearbox. There are some important aerodynamic aspects that have a specific relationship with the mechanical turbine. One is the blade geometry and the incident wind angle of attack. Wind velocity and blade rotating speed have direct effects in the obtained C_p. In order to study these characteristics it is common to construct performance graphics. With this objective in mind, the tip speed ratio coefficient λ is defined [8].

$$\lambda = \frac{r\omega_{tur}}{v} \tag{13.2}$$

where r (rad) is the rotational turbine radius, ω_{tur} (rad/s) is the angular velocity of the mechanical turbine and v is wind speed.

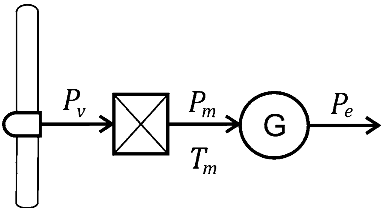


Fig. 13.1 Wind turbine model diagram. Mechanical wind turbine (*left*), gearbox (*middle*), squirrel cage induction generator (*right*)

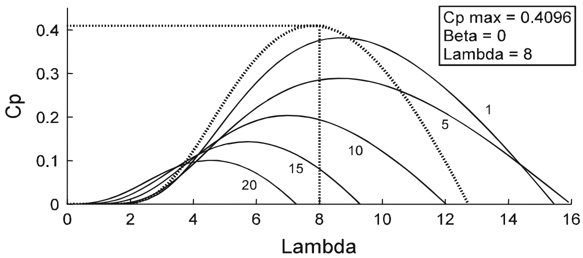


Fig. 13.2 C_p–λ curves for $\beta = 0{,}1{,}5{,}10{,}15$ and 20. *Doted line* corresponds to $\beta = 0$ where $C_{p\ max} = 0.4096$ is achieved

Table 13.1 Wind turbine aerodynamic parameters

Parameter	Value
β	0, 1, 5, 10, 15 and 20
c_1	0.4654
c_2	116
c_3	0.4
c_4	5
c_5	20.24
c_6	0.08
c_7	0.035
λ	0–16

The performance curves commonly used to design a wind turbine for a chosen average site wind speed are the C_p–λ curves. These curves show information regarding wind speed and angle of attack at which maximum power coefficient $C_{p\ max}$ is obtained. The C_p relates with λ with the following expressions [9]:

$$C_p = c_1\left(\frac{c_2}{\lambda_i} - c_3\beta - c_4\right)e^{\frac{-c_5}{\lambda_i}} \tag{13.3}$$

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + c_6\beta} - \frac{c_7}{\beta^3 + 1} \tag{13.4}$$

where c_1, c_2, … c_7 are specific constants for each wind turbine aerodynamic design. β (deg) is the wind angle of attack at the blade.

Figure 13.2 shows C_p–λ curves for different β of the studied 1.5 MW wind turbine.

In Table 13.1 aerodynamic design constants can be found, as well as parameters needed for the drawing of Fig. 13.2 curves.

13.2.2 Gearbox

The gearbox is the mechanical element that multiplies rotational speed of the mechanical turbine ω_{tur} into the speed needed for the electric generator ω_m. This generation rotational speed is generally slightly faster than the synchronous speed

ω_s. For the Mexican grid that works at a 60 Hz frequency, $\omega_s = 2\pi(60\text{ Hz}) = 376.99\text{ rad} \approx 377\text{ rad}$. Thus, electric generator rotational speed is:

$$\omega_m = n\omega_{tur} \tag{13.5}$$

where n is a multiplicative factor and ω_{tur} is the mechanical turbine (*low speed rotor*) angular velocity.

The mechanical power P_m delivered at the output of an ideal gearbox is the same as the one extracted from wind and multiplied by the power coefficient C_p, $P_m = C_p(\beta,\ \lambda)P_v$. For wind at standard conditions (101.3 kPa *y* 273 K) density value is $\rho = 0.647$ (kg/m^3), thus:

$$P_m = 0.647 C_p(\beta, \lambda)\frac{1}{2}Av^3 \tag{13.6}$$

This mechanical power (W) is transmitted to the electrical generator with the following expression of mechanical torque.

$$T_m = \frac{P_m}{\omega_m} \tag{13.7}$$

where T_m is mechanical torque and ω_m is angular speed, both at the fast rotational side of the gearbox (*rotational speed of the electrical generator*).

13.2.3 Electrical Generator

For the simulation, a squirrel cage induction generator was selected given that this type of generator is the most commonly used.

The squirrel cage induction generator model (*Asynchronous machine*) was obtained from Simulink MatLab®.

The mechanical turbine inertia constant was added with the electrical generator own inertia, taking into account that this constant is generally ten times bigger in comparison with the generators' [10].

13.2.4 Implementation of Wind Turbine Model

The complete wind turbine model was implemented in Simulink of MatLab®. The mechanical turbine was constructed with Eqs. (13.3), (13.4) and (13.6). The gearbox was modeled as a simple speed gain as in Eq. (13.5). Input to the implemented model is wind speed incident to the mechanical turbine. Model's output is generated electrical power by the squirrel cage induction generator model. Parameters for the 1.5 MW wind turbine can be found in Table 13.2.

13.3 Control Design

The most common way of controlling a wind turbine consists in varying attack angle β at the blades (*pitch control*) in order to modify the mechanical turbine aerodynamic characteristics and thus modify its performance in accordance to changing wind speed. The blade can be pitched with two methodologies: pitching to stall or pitching to feather. The selection of one or other method has important effects in the wind turbine aerodynamic characteristics. With the pitching to stall method, pitch control is achieved with small negative angle adjustments. The problematic with this methodology is due to undesirable damping and fatigue effects that cannot be effectively modeled. Pitching to feather is the preferred methodology due to the form the wind surrounds the blade. This aerodynamic effect can be easily modeled and as a consequence, mechanical stress can be foreseen with more reliability. The problem with this type of pitch control is that much bigger β angles are needed to effectively control the wind turbine, in this case positive [7]. For the present work, pitching to feather methodology was chosen as can be observed in the positive β angles in Table 13.1.

13.3.1 Proportional Integral Controller

A Proportional-Integral (PI) control is a special case of the classic controller family known as Proportional-Integral-Derivative (PID). These type of controllers are up to

Table 13.2 Wind turbine model parameters

Mechanical turbine	
Parameter	Value
r	34 m
A	πr^2
Gearbox	
Parameter	Value
n	152.49
Generator	
Parameter	Value
P_{nom}	1.5 MW
V_{nom}	575 V
F_{nom}	60 Hz
R_s	0.004843 pu
L_{ls}	0.1248 pu
R_r	0.004377 pu
L_{lr}	0.1791 pu
L_m	6.77 pu
H	$H_{tur} + H_g = 4.125$ s
F	0.01 pu
poles	3

date the most common way of controlling industry processes in a feedback configuration. More than 95 % of all installed controllers are PID [11, 12].

For the designed PI controller for the 1.5 MW horizontal axis wind turbine error signal was selected as:

$$e(t) = P_{ed}(t) - P_e(t) \tag{13.8}$$

where P_{ed} is the desired output or set point for the wind turbine, in this case 1.5 MW, and P_e is the actual delivered power from the wind turbine [7].

The PI control was optimized to achieve rapid response to different wind speed changes and to deliver nominal power output for nominal wind speed (11.75 m/s) as well as higher wind speeds. An open loop analysis was performed at an operating mode for nominal wind speed and without altered aerodynamic blade pitch conditions ($v = 11.75$ m/s, $\beta = 0°$). The Ziegler–Nichols tuning method was then applied to obtain initial gains, which were modified on a trial and error basis to obtain a desirable response. Obtained gains were: $K_p = -0.934$ and $T_i = 0.444$. Due to the big input signal to the controller, the error was divided by a 10^5 factor and the integral action saturated at a -45 lower level and a 0 upper level.

13.3.2 Fuzzy Logic Controller

A Fuzzy Logic Controller (FLC) is basically designed by selecting its inputs and outputs, choosing the preprocessing needed for the inputs and the post-processing needed for the outputs, as well as designing each of its four basic components: *fuzification*, *rule-base*, *inference mechanism* and *defuzification*. An FLC is an artificial decision making system that operates in closed loop and real time as can be observed in Fig. 13.3. A more detailed explanation of this methodology can be found in [6, 13] and [14].

For the proposed FLC, inputs to the controller are wind speed $v(t)$ and an $e(t)$ signal as the one in Eq. (13.8). The closed loop diagram for the proposed FLC is shown in Fig. 13.3.

In order to understand the way electric power from the wind turbine is obtained using the pitching to feather methodology, performance curves can be constructed (like in Fig. 13.2). The most useful performance curves for this purpose are the

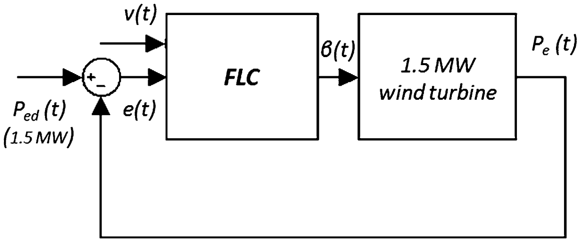


Fig. 13.3 Feedback control closed loop for the FLC. Showing inputs and control output

P_e–v, which show generated electric power versus wind velocity at constant chosen β angles. Figure 13.4 shows some of these curves.

The P_e–v curves were drawn for chosen $\beta = 0$, 2, 12, 18 and 23. From the curves it is obvious how β angles should be increased in order to maintain a 1.5 MW power generation. Additionally to nominal value power extraction it is also important to obtain maximum generation at low wind speeds.

From Fig. 13.4 we can observe that for wind speeds below 8 m/s, ideal angle for maximum power extraction is $\beta = 2°$. This is an interesting fact because most fixed speed wind turbines maintain a $\beta = 0°$ for speeds below the nominal wind speed. Rules can be derived from observation of these performance curves. Rules loaded to the designed FLC can be found in Table 13.3.

Inference mechanism is basically defined with membership functions, which are used to determine the relevance of the set of rules of Table 13.3. Implemented membership functions are shown in Figs. 13.5, 13.6 and 13.7, $v(t)$ and $e(t)$ inputs and $\beta(t)$ output respectively. Methods for implication and aggregation where defined as *minimum* and *maximum* respectively. Defuzification process was selected as centroid [5].

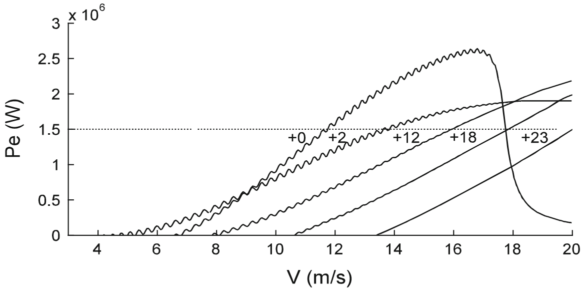


Fig. 13.4 P_e–v curves for $\beta = 0$, 2, 12, 18 and 23. Optimum β angle for current wind speed can be obtained at the intersection with the 1.5 MW *doted line*

Table 13.3 FLC set of rules

v (m/s)	Power e(t)				
	NegVB	NegB	Accept	PosB	PosVB
5	0	1	2	2	2
7	0	1	2	2	2
9	2	2	1	1	0
11	1	0	0	0	0
11.7	1	0	0	0	0
12.6	6	2	1	0	0
13.8	10	6	2	1	0
14.8	14	10	6	2	1
15.5	18	14	10	6	2
16.5	20	18	14	10	6
17.8	20	20	18	14	10
18.6	22	22	20	18	14
19.5	24	24	22	20	18
20.5	24	24	24	22	20

13.3.3 Fuzzy Proportional Integral Controller

The designed Fuzzy Proportional Integral (Fuzzy-PI) controller is a hybrid controller that utilizes two sets of PI gains in order to achieve a non-linear response. The switching in this controller is achieved with a fuzzy logic section that depends on the input *v*(*t*). The PI gains utilize *e*(*t*) as in Eq. (13.8). Figure 13.8 shows a diagram of the proposed Fuzzy-PI controller.

For the proposed Fuzzy-PI algorithm, the fuzzy logic section was designed to smoothly switch between low speed PI gains and nominal and faster wind speed PI gains. The switching was performed following a heuristic approach based in analyzing optimum β angles for different wind speeds from Fig. 13.4. Following this reasoning the appropriate set of rules was constructed. These rules can be found on Table 13.4.

From Table 13.4, left column is fuzzy input wind speed *v*, with two levels, LWS, low wind speed and FWS, high wind speed. Columns to the right are fuzzy outputs K_p and T_i, which have two levels as well. LSKp, low speed K_p, FSKp, fast speed K_p, LSTi, low speed *Ti*, FSTi, fast speed T_i.

In order to obtain the appropriate PI gains for nominal (11.75 m/s) wind speed and lower than nominal wind speeds the same methodology as in PI control section was used. For lower than nominal wind speed tuning was achieved with the same

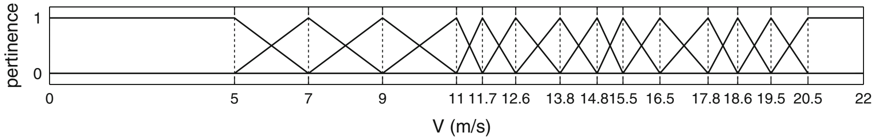


Fig. 13.5 v(t) FLC input membership functions

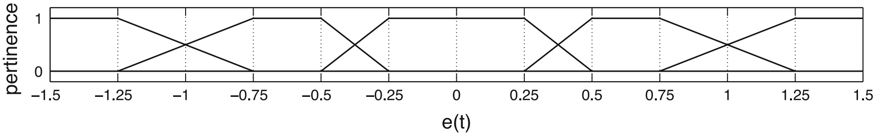


Fig. 13.6 *e*(*t*) FLC input membersip functions. NegVB, NegB, accept, PosB and PosVB from *left* to *right*

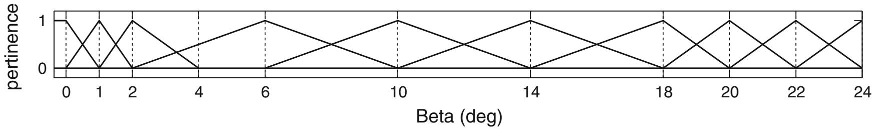


Fig. 13.7 $\beta(t)$ FLC output membersip functions

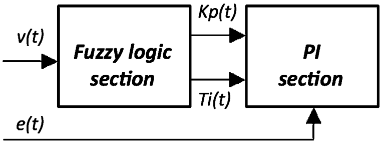


Fig. 13.8 Fuzzy PI controller diagram. Input to fuzzy logic section is wind speed, outputs are PI gains. Input to PI section is error signal, output is control $\beta(t)$

Table 13.4 Fuzzy logic section set of rules

v	K_p	T_i
LWS	$\mathrm{LSK_p}$	$\mathrm{LST_i}$
FWS	$\mathrm{FSK_p}$	$\mathrm{FST_i}$

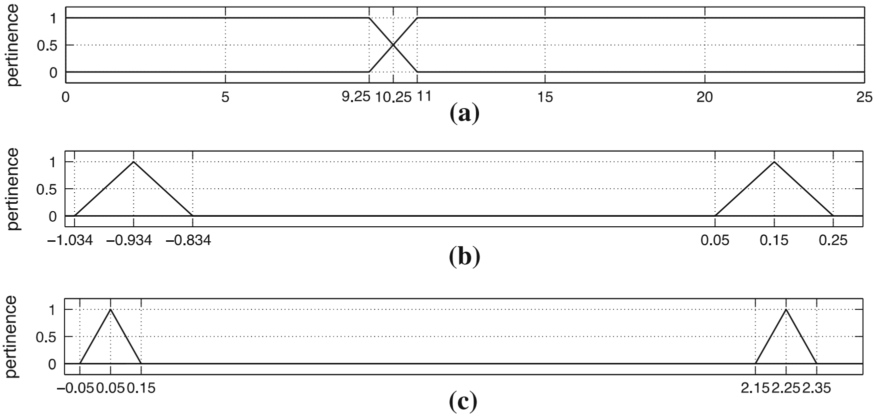


Fig. 13.9 **a** Fuzzy logic section input wind speed $v(t)$. LWS fuzzy set at *left* and FWS fuzzy set at *right*. **b** Fuzzy logic section output proportional gain K_p. FSKp fuzzy set at *left* and LSKp fuzzy set at *right*. **c** Fuzzy logic section output integral gain $1/Ti$. LSTi fuzzy set at *left* and FSTi fuzzy set at *right*

1.5 MW set point ($v = 6$ m/s, $\beta = 2$). Obtained gains for low wind speed operation were: $K_p = 0.15$ and $T_i = 20$.

Implemented membership functions are shown in Fig. 13.9a–c, $v(t)$ input, $K_p(t)$ and $T_i(t)$ outputs respectively. Methods for implication, aggregation and defuzification process where defined as in the FLC.

13.4 Simulation and Results

For simulation purposes, wind signal was constructed with two different profiles for a 300 s period. Figures 13.12 and 13.13 show these wind profiles, which correspond to near nominal and faster wind speed operation and low speed operation respectively. For the robustness test, noise was added to the wind profiles in Fig. 13.10. Figure 13.11 shows noisy wind profiles.

Control responses were obtained for both wind speed operation signals (Fig. 13.10). Figures 13.12 and 13.13 show power extraction and control signal results.

From Fig. 13.12 top, it can be seen how all control algorithms obtain adequate power control at nominal 1.5 MW level. However, low $\beta(t)$ angles (Fig. 13.12 bottom) present at the FLC and Fuzzy-PI algorithms achieve better performance for low wind speeds.

Figure 13.13 top, clearly shows how FLC and Fuzzy PI algorithms performance surpass PI control algorithm performance for low wind speeds. Maximum power extractions for FLC and Fuzzy-PI are due to small $\beta(t)$ angles near 2° as can be observed from Fig. 13.13 bottom.

Figures 13.14 and 13.15 show results for extracted power and control signal for noisy wind speed input (Fig. 13.11) to controllers for FLC and Fuzzy-PI methodologies.

From Fig. 13.14 top, it can be seen how FLC has no adequate power control at nominal 1.5 MW level because of its dependence on wind speed input to operate. However, Fuzzy-PI algorithm responds in an adequate form. This can also be observed at Fig. 13.14 bottom, where control signal for the FLC methodology is affected by noise.

These same results can be observed in Fig. 13.15 bottom, for lower than nominal wind speeds. However, Fig. 13.15 top, shows no clear advantage between FLC and Fuzzy-PI controller.

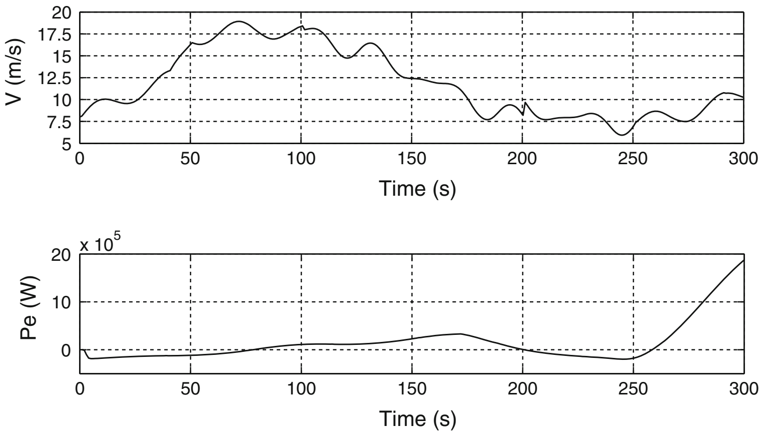


Fig. 13.10 Near nominal wind speed operation $v(t)$ signal, *top*. Low wind speed operation $v(t)$ signal, *bottom*

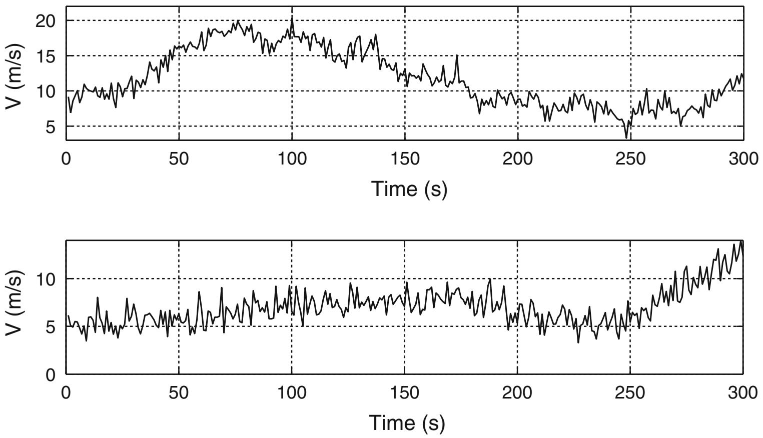


Fig. 13.11 Near nominal wind speed operation, noisy *v*(*t*) signal, *top*. Low wind speed operation, noisy *v*(*t*) signal, *bottom*

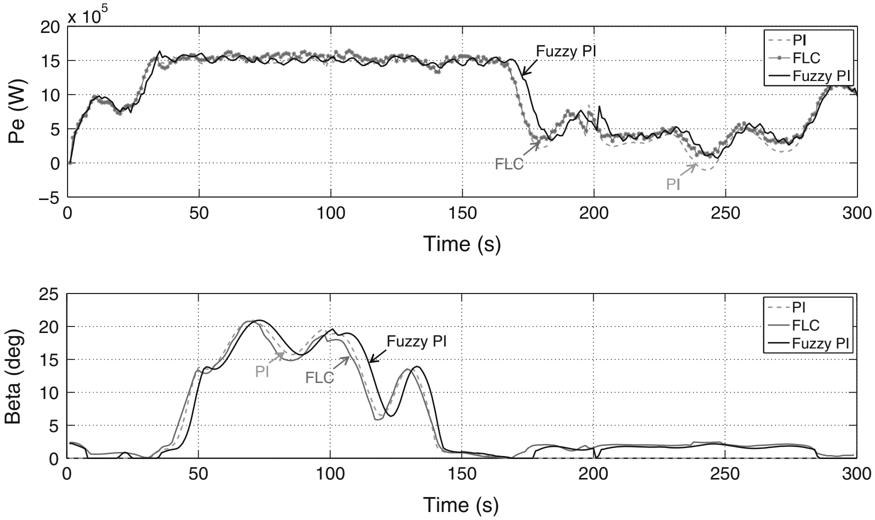


Fig. 13.12 P_e, *top*. $\beta(t)$ control output, *bottom*. Near nominal wind speed operation

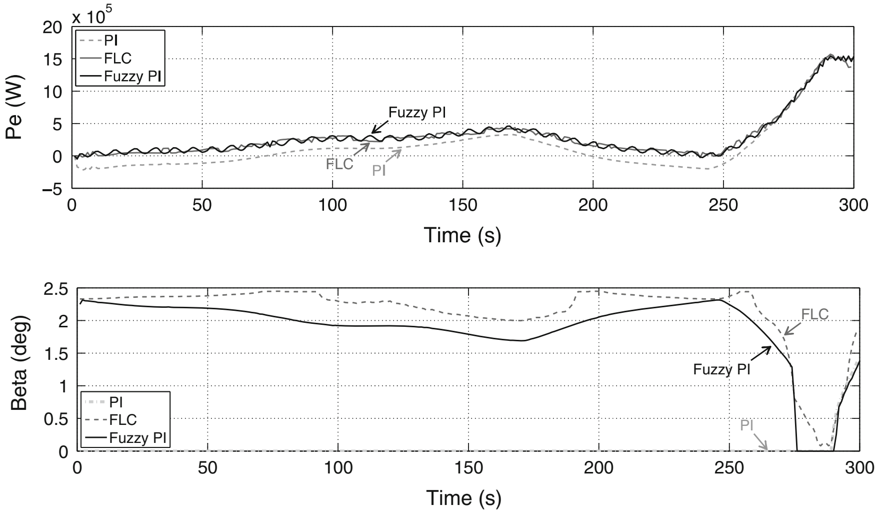


Fig. 13.13 P_e, *top*. $\beta(t)$ control output, *bottom*. Low wind speed operation

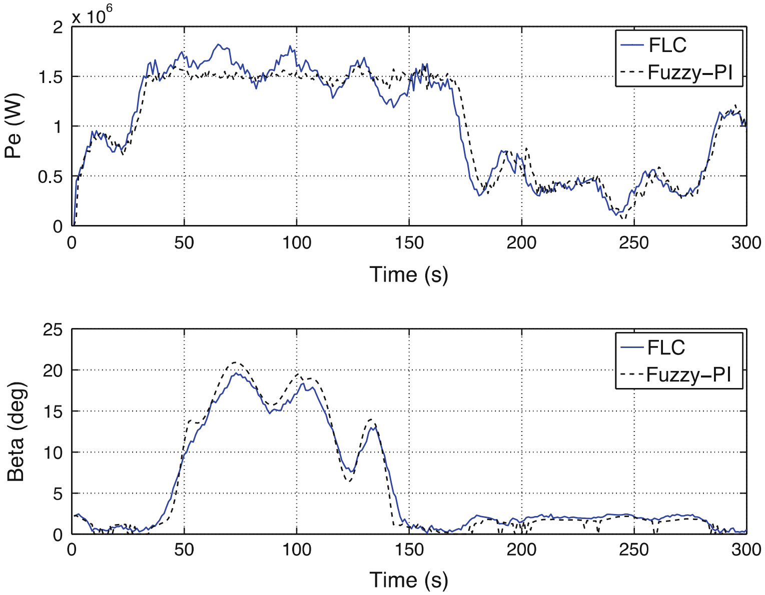


Fig. 13.14 P_e, *top*. $\beta(t)$ control output, *bottom*. Noisy near nominal wind speed operation

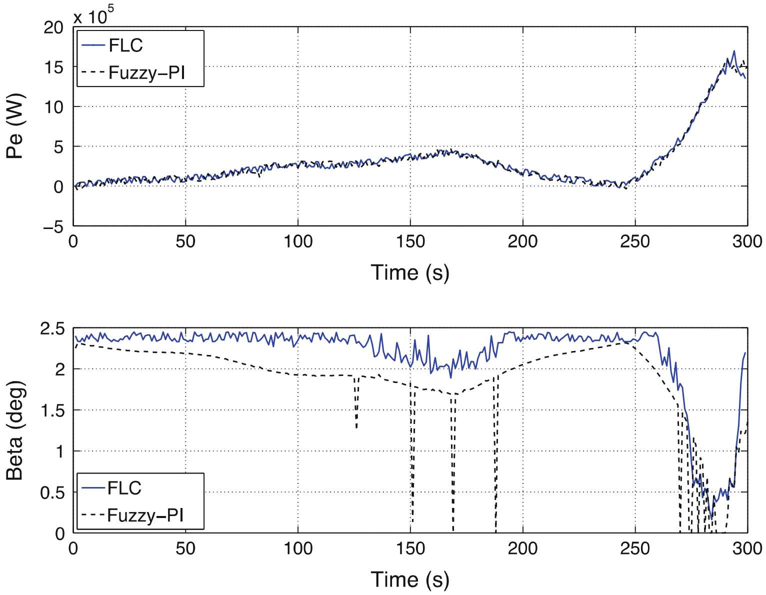


Fig. 13.15 P_e, *top*. $\beta(t)$ control output, *bottom*. Noisy low wind speed operation

13.5 Conclusion

Research presents a PI, FLC and PI control comparison in simulation for a 1.5 MW horizontal axis fixed speed wind turbine model, PI control algorithm achieves good performance for power extraction near the nominal 1.5 MW for nominal wind speeds (around 11.75 m/s) and higher speeds, however, a constant 0° $\beta(t)$ angle at lower speeds results in poor power extraction. The implemented FLC and Fuzzy-PI control algorithms surpass the traditional PI thanks to its inherent characteristics to deal directly with non linear models. From quick inspection to power versus wind extraction performance curves all needed control rules can be extracted. A fuzzy logic section in the Fuzzy-PI algorithm allows for a non linear operation using a smooth PI gain switching methodology with good results. A robustness test was performed by adding a noisy wind speed signal to the FLC and Fuzzy-PI control algorithms. Results demonstrate the inherent ability of the Fuzzy-PI control algorithm to deal with this kind of noise. FLC algorithm shows no adequate response for noisy $v(t)$ signal, which is understandable because of its dependence in this control input. These results are important because in a real scenario wind speed measurement is a difficult task.

References

1. Barroso LA, Rudnick H, Sensfuss F, Linares P (2010) The green effect. IEEE Power Energy, IEEE PES, EUA, 8(5):22–35
2. Anzurez-Marin J, Torres-Salomao LA, Lázaro-Castillo II (2011) Fuzzy logic control for a two tanks hydraulic system model. In: Proceedings 2011 IEEE electronics, robotics and automotive mechanics conference CERMA 2011, Cuernavaca, Mexico
3. Torres-Salomao LA, Gámes-Cuatzin H, Anzurez-Marín J, Lázaro-Castillo II (2012) Fuzzy-PI control, PI control and fuzzy logic control comparison applied to a fixed speed horizontal axis 1.5 MW wind turbine. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS 2012, San Francisco, USA, 24–26 Oct 2012, pp 1181–1186
4. Hong SK, Nam Y (2003) An LMI-based fuzzy sate feedback control with multi-objectives. KSME Int J, Springer, Korea, pp 105–113
5. Torres-Salomao LA, Gámez-Cuatzin H (2012) Fuzzy logic control and PI control comparison for a 1.5 MW horizontal axis wind turbine. In: 16th International conference on system theory, control and computing, ICSTCC, Control Society, Sinaia, Romania, pp 1–6
6. Passino KM, Yurkovich S (1998) Fuzzy control. Addison-Wesley, USA
7. Burton T, Sharpe D, Jenkins N, Bossanyi E (2001) Wind energy handbook. Wiley, England
8. Johnson GL (2006) Wind energy systems. Electronic edition, USA
9. Kyoungsoo R, Choi H (2004) Application of neural network controller for maximum power extraction of a grid-connected wind turbine. Springer
10. Saad-Saoud Z, Jenkins N (1995) Simple wind farm dynamic model. IEEE Proc Gener Transm Distrib 142(5):545–548
11. Åström KJ, Hägglund TH (1995) New tuning methods for PID controllers. In: Proceedings of the 3rd European control conference, pp 2456–2462
12. Lázaro-Castillo II (2008) Ingeniería de Sistemas de Control Continuo. UMSNH, COECyT Michoacán, FIE, Mexico
13. Heske T, Neporent J (1996) Fuzzy logic for real world design. Annabooks, San Diego
14. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. Inf Sci 8:199–249

Chapter 14
Linear Quadratic Regulation of a Rotating Shaft Being Subject to Gyroscopic Effect Using a Genetic Optimization Algorithm

Rudolf Sebastian Schittenhelm, Matthias Borsdorf, Zhentao Wang and Stephan Rinderknecht

Abstract A Linear Quadratic Regulator and a Kalman Filter are designed for a rotor test rig being subject to unbalance excitation and gyroscopic effect. Rotor vibration is controlled by means of two piezoelectric stack actuators installed at one of the two supports of the rotor. The presence of gyroscopic effect leads to an undesirable dependence of the system dynamics on rotational frequency of the shaft. As a result, there is a need for high robustness and furthermore, the separation principle does not hold. Due to the latter aspect, controller and observer design become a coupled problem in the case of the rig. In a first step, the number of free design parameters of the controller-observer combination is reduced to a manageable number of 5. Subsequently, these parameters are determined by means of a genetic optimization algorithm on the basis of a Finite Element model of the test rig. It is shown, that it is possible in this way to determine a controller-observer combination leading to robust stability and excellent performance in the whole operating range which contains two unbalance induced resonances. Control performance is validated in simulation as well as experiments at the test rig.

Keywords Active vibration control · Genetic optimization algorithm · Gyroscopic effect · Kalman filter · Linear quadratic regulator · Piezoelectric actuator · Rotor vibration

R. S. Schittenhelm (✉) · M. Borsdorf · Z. Wang · S. Rinderknecht
Institute for Mechatronic Systems in Mechanical Engineering, Petersenstr. 30
64287 Darmstadt, Germany
e-mail: schittenhelm@ims.tu-darmstadt.de

M. Borsdorf
e-mail: borsdorf@ims.tu-darmstadt.de

Z. Wang
e-mail: wang@ims.tu-darmstadt.de

S. Rinderknecht
e-mail: rinderknecht@ims.tu-darmstadt.de

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_14,

14.1 Introduction

Vibration is a critical aspect in high speed rotor applications such as e.g. aircraft engines or high speed cutting machines. Vibration related matters gain cumulatively in importance since the demand for lightweight construction leads to lighter machines being prone to dynamic excitation. In many cases, attenuation of vibration amplitudes can be achieved by means of passive measures, i.e. balancing, damping elements like squeeze film dampers (SFD) or targeted manipulation of eigenfrequencies and eigenvectors. As those methods reach their limits in certain applications, active means are an attractive alternative. Furthermore, active measures for vibration control are more flexible in terms of the objective to be achieved. For instance, different controllers can be used for different operating conditions using the same hardware. In this way it is possible to switch between vibration control and vibration isolation for example, which is not possible by passive means to the same extent.

Possible (semi-) active components for use in vibration control of rotating machinery are active SFDs [1], electrorheological dampers [2], active magnetic bearings and electromagnetic actuators [3–7] as well as piezo-actuated bearings [8–11]. Advantages of piezoelectric stack actuators are low weight accompanied by high forces in a broad frequency range if a suitable amplifier is available. Furthermore, even though there is hysteresis present in piezoelectric material, nonlinearity is relatively weak and as a result, a linear approximation of the piezoelectric effect is admissible for controller design purposes.

The heart of any control system is the controller itself. Feedforward as well as feedback control is frequently applied to vibration control problems throughout literature. Both approaches are briefly discussed in the following.

Feedforward methods are commonly adaptive because of inevitable model mismatch and a lack of knowledge about the disturbance. The algorithm most frequently used for the adaption process in the context of vibration control is the Filtered x Least Mean Squares (FxLMS) algorithm [3, 9, 12]. In [3, 9], the suitability of the FxLMS algorithm for vibration control of rotating machinery is investigated and high vibration attenuation is achieved in these articles. In contrast to model based feedback controller design, no model of the frequency response functions (FRFs) from disturbances to outputs is required for implementation. In most rotor applications, these FRFs cannot be fully determined due to a lack of knowledge about the disturbance location, such that this aspect is considered to be a great advantage. However, the convergence behavior of the FxLMS algorithm is hard to predict under fluctuating operating conditions. This is especially true for systems having time dependent dynamics, such as the system treated in this chapter, since the algorithm requires models of the so-called secondary path FRFs from the actuators to the sensors. If there is a phase mismatch in the secondary path model of 90° or more, the algorithm diverges in the single input single output case [12]. This is critical for weakly damped mechanical systems, since there are steep phase jumps of 180° at resonance as well as anti-resonance frequencies.

Thus, just small mismatch of resonance or anti-resonance frequencies in the model leads to large mismatch in secondary path phase, which makes implementation of the FxLMS difficult.

Feedback methods for vibration control, from the authors' point of view, can be classified into three categories:

Classical controllers such as PID-control, Integral Force Feedback, Positive Position Feedback and similar techniques, see [13], build the first category. They are frequently used for academic as well as practical investigations and can be implemented without a model of the system under consideration. The often time-consuming modeling process can thus be skipped if one decides to use this non-model based approach for controller design. However, performance of classical controllers is often suboptimal due to a lack of closed solutions for calculation of optimal controller gains of multiple input multiple output systems. Classical controllers are usually designed in a decentralized manner, i.e. just one output is fed back into one input. As a result, the potential of off-diagonal entries of the controller FRF matrix for vibration control is not utilized.

The basic idea of frequency domain approaches, e.g. μ-, H_∞- or H_2-optimal controllers, is a more intuitive controller design by means of performance specification via frequency dependent weighting functions for the FRFs of the system. Inclusion of model inaccuracy in the controller design process is possible when using μ- or H_∞-optimal controllers. Thus, a possible demand for robustness can be explicitly considered in the controller design process if an uncertainty model is available. The problem description in the frequency domain can be considered to be convenient for mechanical engineers involved in structural dynamics, since they usually are familiar with system description via FRFs. However, the definition process of weighting functions offers infinite degrees of freedom as these functions can be shaped arbitrarily. As a result, weighting functions usually have to be defined in a highly iterative manner in order to achieve a well-performing controller. Furthermore, and in contrast to state space controllers, the order of frequency domain controllers usually exceeds the model order. The order of H_∞-optimal controllers is given by the order of the weighted plant model [14], while the order of μ-optimal controllers is usually even higher. Thus, controller reduction is required after controller design if there are limits regarding computational effort of the real time system. Controller reduction may reduce the closed loop performance significantly and may even lead to unstable closed loop systems. In rotordynamics, frequency domain controllers are commonly used in the context of active magnetic bearings, see [4, 5] for example.

State space controllers form the third category. The classic examples, i.e. the Linear Quadratic Regulator (LQR) and Pole Placement controllers, mainly aim at the alteration of system dynamics and do not account for disturbance rejection behavior. A major drawback of state space controllers is the need for observers if not all system states are measurable, which is usually the case for rotor systems. This aspect is particularly challenging if the system dynamics vary during operation, leading to the fact that separation principle does not hold if the classic state space controller approaches and a nominal model at some specific design point are utilized for observer design. Also, disturbances are not considered in classic observer design, leading to an inaccurate state estimate if unbalance excitation is present. However, classic state space controllers and observers offer the user simple implementation and a good understanding of the alteration of the dynamic characteristics of the system by the controller. Also, there are advanced observer techniques available for rotors excited by unbalance and subject to gyroscopic effect [15], which consider system deviation as well as disturbances in the form of unbalances. State space controllers are thus frequently applied to control problems throughout literature. The LQR possesses the desirable property of being robust to modeling errors in the systems input matrix [16]. The Kalman Filter is the counterpart observer to the LQR. It can be designed by applying the LQR procedure to the dual system with a certain weighting and can be tuned to be insensitive to measurement or actuator noise [16]. For these reasons and due to their ease of implementation, LQRs and Kalman Filters are the most frequently used state space controller and observer types in the field of active vibration control of rotor systems. The LQR has been successfully applied to rotor systems excited by unbalance for example in [6, 7, 10]. Results indicate that the controller is suitable for the problem since good vibration attenuation is achieved. However, even though dependence of the system dynamics on the rotational frequency due to gyroscopic effects is mentioned in some of these papers [6, 10], the problem of model mismatch is not addressed in detail.

In this chapter, a LQR and a Kalman Filter, resulting in a Linear Quadratic Gaussian (LQG) controller [16], are applied to a rotor system. The choice of these types of controller and observer is due to their robustness, ease of implementation, noise insensitivity and the advantage of the design parameters being physically interpretable. The problem of model mismatch due to gyroscopic effect is treated and overcome.

This investigation is an extension of the results published in [11], where a proposal for efficient brute force optimization or manual tuning of 3 design parameters was given. In this article however, a controller structure having 5 free design parameters is proposed and the controller parameters are found fully automatically using a genetic optimization algorithm [17–20], which is readily available in the MATLAB environment [19].

Genetic algorithms (GA) aim at emulating the process of natural evolution in a population. By applying the rules of evolution, namely selection, crossover and mutation [20], to a population, i.e. a set of candidate solutions to the optimization problem, the algorithm is assumed to approach better and better solutions with each new generation of individuals created. Performance of GAs can be enhanced by means of combination with other, e.g. gradient based optimization algorithms, leading to an improvement of convergence [18, 20] by means of better exploitation of the regions of attraction of good solutions [20]. This combination is referred to as a hybrid GA [18]. It is not taken advantage of in this article for the sake of brevity.

By means of the proposed controller structure and a GA, robust stability and very high vibration attenuation is achieved in the whole operating range of the rotor, which covers two unbalance induced resonances. The controller is validated in simulation and experiment.

14.2 System Description and Modeling

The controllers designed in this chapter are applied to the rotor test rig shown in Fig. 14.1. The rotating shaft is of approximately 9 mm diameter and 320 mm length and has two discs mounted on it. Mainly due to the cantilever disc 1, the rotor is subject to gyroscopic effect to significant extent, as intended by the authors in order to replicate a close to reality high speed rotor application. There are two piezoelectric actuators installed in the active bearing in order to influence rotor vibration. They possess a maximum stroke of 60 μm at the maximum admissible voltage of 1,000 V. The actuators receive a voltage offset of 500 V during operation, such that the maximum admissible voltage amplitude is $u_{\max} = 500\,\mathrm{V}$. Power supply is achieved by two amplifiers with a maximum admissible current of $\pm 100\,\mathrm{mA}$. Two springs are mounted on the opposite side of the rotor in order to apply pre-stress to the actuators. Besides the active bearing, which is located between the discs and is shown in Fig. 14.2 in more detail, there is a passive bearing at the other end of the shaft.

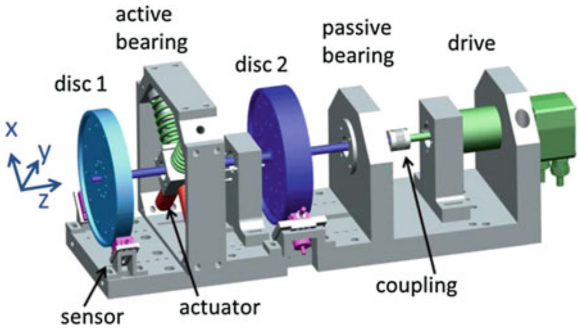


Fig. 14.1 Test rig

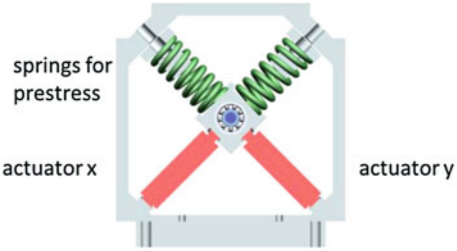


Fig. 14.2 Active bearing

Displacements of the discs into the x and y direction are measured by four eddy current sensors. The resulting sensor signals are low pass filtered by means of a 1 kHz first order analogue low pass filter in order to avoid aliasing effects in the digitalization process of the measurement data. The rotor is accelerated by means of a 250 W DC motor. The maximum rotational frequency of the rotor is 160 Hz. It has two unbalance induced resonance frequencies in its operating range at approximately 47 and 108 Hz. A dSpace real time system (DS1104) is utilized for data acquisition as well as controller implementation and run with a sampling frequency of 2.5 kHz.

A model of the rig is derived by means of Finite Element (FE) analysis on the basis of Timoshenko beam theory using a FE program written at the Institute for Mechatronic Systems in Mechanical Engineering of the Technische Universität Darmstadt in the MATLAB environment. A FE model is used rather than an identified one in order to achieve a model in which the states are physically interpretable, i.e. assignability to vibration modes in this case. Furthermore, the gyroscopic matrix is difficult to identify accurately, whereas it can easily be derived using FE method. Bearings are modeled by means of discrete spring, mass and piezoelectric elements. The piezoelectric effect is incorporated using a linear description, even though in reality there is some hysteresis present [10, 11, 13]. The results achieved in literature [10, 11] as well as this chapter using a linear description show, that this simplification is permissible for controller design purposes. Damping is introduced by means of viscous modal damping and damping ratios of 0.6–1 % which were tuned together with other system parameters by means of a model updating routine. The model is reduced by means of modal reduction technique to an order of 16 in state space, i.e. 8 modes. The rotor system, including piezoelectric actuators is described by

$$\begin{aligned} \dot{x} &= A_\Omega x + B_\Omega u + E_\Omega d \\ y &= C_\Omega x + D_\Omega u. \end{aligned} \tag{14.1}$$

$x \in \mathbb{R}^n$ denotes the vector of system states, $u \in \mathbb{R}^{n_u}$ the control inputs, i.e. the voltages applied to the actuators, $d \in \mathbb{R}^{n_d}$ the disturbances and $y \in \mathbb{R}^{n_y}$ the sensor signals. A_Ω, B_Ω, C_Ω, D_Ω, E_Ω are the respective system matrices with appropriate dimensions. The states in the system (14.1) are arranged with ascending eigenfrequencies. Due to gyroscopic effect, the system dynamics depend on the rotational frequency $f = \Omega/2\pi$, which is indicated by the respective subscript of the system matrices in (14.1). The reduced model is augmented in order to capture the effects of filters, amplifiers and digital signal processing. Low pass filters and amplifiers are described by means of first and second order low pass models respectively and the effect of digital signal processing is introduced to the model by a second order padé approximation of a time delay of one sample. For the purposes in this chapter, it is assumed to be admissible to add the filters to the model at the inputs rather than on the physically correct outputs. Thus, effects of filters, amplifiers and sampling are described by a single 2 × 2 model of order 10,

$$\begin{aligned} \dot{x}_P &= A_\mathrm{P} x_P + B_\mathrm{P} u_{in} \\ u &= C_\mathrm{P} x_P, \end{aligned} \tag{14.2}$$

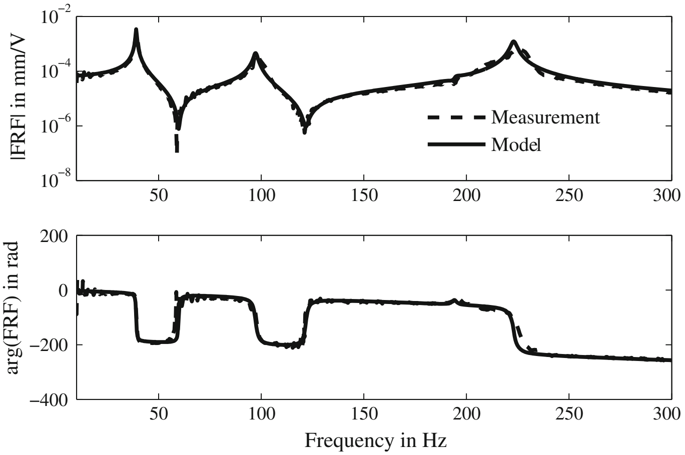


Fig. 14.3 FRF of plant and model from actuator x to the sensor at disc 1 pointing into the x-direction

and included into the overall model as shown below:

$$\begin{aligned} \dot{x}_S &:= \begin{bmatrix} \dot{x} \\ \dot{x}_P \end{bmatrix} = \begin{bmatrix} A_\Omega & B_\Omega C_{\mathrm{P}} \\ 0 & A_{\mathrm{P}} \end{bmatrix} \begin{bmatrix} x \\ x_P \end{bmatrix} + \begin{bmatrix} E_\Omega & 0 \\ 0 & B_{\mathrm{P}} \end{bmatrix} \begin{bmatrix} d \\ u_{in} \end{bmatrix} \\ &=: A_{S,\Omega} x_S + E_{S,\Omega} d + B_S u_{in} \\ y &= [\, C_\Omega \quad D_\Omega C_P \,] x_S =: C_S x_S \end{aligned} \tag{14.3}$$

The model matches reality with high accuracy, see Fig. 14.3, where the modeled and identified FRFs from the actuator pointing into the x direction to the sensor pointing into the x direction at disc 1 are shown for the nonrotating rotor ($f = 0\,\mathrm{Hz}$). Effects of low pass filters, amplifiers and sampling are noticeable when inspecting the phase drift of the transfer functions in Fig. 14.3. It is assumed that controller implementation is not possible in a targeted manner without consideration of these elements.

For model based controller design, it is necessary to obtain an approximation of the disturbance. As in [11], the unbalance excitation for simulation and controller design is identified by means of analysis of measurement data at the critical speeds. A method in the frequency domain being similar to influence coefficient balancing method [21] is applied. For more details, see [11].

14.3 Controller Structure

Controller design is challenging in the context of the rotor under consideration because of the dependence of the system dynamics (14.3) on the rotational frequency. Due to fluctuation of system characteristics, there is a demand for high robustness. Furthermore, the separation principle does not hold if a simple constant observer is applied. In this section, the coupled problem of LQR and Kalman Filter design for the rig is treated and a proposal to reduce the problem to 5 design parameters is given.

Linear Quadratic Regulators minimize the objective function J_1,

$$J_1 = \int_0^{\infty} \left(x_s^T Q x_s + u_{in}^T R u_{in}\right) dt, \tag{14.4}$$

for arbitrary $x_s(0) \neq 0$ and $d \equiv 0$. Q and R are weighting matrices, which are chosen to be of diagonal shape in this chapter.

Obviously, the objective function (14.4) does not accurately describe the problem of a rotor being subject to unbalance excitation since it is rather a measure of settling time and no disturbances are considered. However, the damping ratios ζ_i of modes of interest can be increased by proper weighting, leading to reduced amplitudes at resonance. The weighting matrix for the control input is defined to be $R = I_2$, where $I_k \in \mathbb{R}^{kxk}$ denotes the identity matrix. Q is chosen to be of the shape

$$Q = diag(10^{q_3} I_2,\ 10^{q_1} I_2,\ 10^{q_3} I_2,\ 10^{q_2} I_2,\ 10^{q_3} I_8,\ 0\,I_{10}). \tag{14.5}$$

In this way, the modes corresponding to the unbalance induced resonances are assigned the weights 10^{q_1} (for the first forward whirl mode) and 10^{q_2} (for the second forward whirl mode) respectively. All other rotor modes are weighted by 10^{q_3} and the modes corresponding to the model of filters, amplifiers and sampling are not weighted, which is achieved by the last block in (14.5). Due to the fact, that the system states in (14.3) are not entirely measurable in the case of the test rig, an observer is required in order to achieve state feedback. The observer feedback matrix, denoted by L, is computed by applying the LQR design method to the dual system, leading to a so-called Kalman filter being insensitive to measurement and actuator noise if designed properly. If one assumes noise of equal level at all sensors and also at all actuators respectively, it is sufficient to choose the weights for the observer as in (14.6):

$$R_o = 10^{r_B} I_4,\ Q_o = B_{\Omega_D} B_{\Omega_D}^T \tag{14.6}$$

r_B describes the ratio of the assumed measurement and actuator noise levels and Ω_D is the design point frequency. The value r_B is not assigned on the basis of noise levels at the rig in this investigation. Nonetheless, this physical consideration helps reducing the number of free parameters to be determined in observer design to a single value.

The matrices of the differential equations of the system are dependent on rotational frequency due to gyroscopic effect, while the controller and observer applied to the system are linear time invariant. Therefore, one has to choose a specific design point frequency $\boldsymbol{\Omega_D}$, which is used for controller and observer design. Due to the facts that the system differs from the nominal behavior at $\boldsymbol{\Omega_D}$ and that neither unbalance excitation nor steady state behavior is considered in the controller objectives, one has to check whether the controller responds adequately to unbalance excitation in the whole operating range $[0; \Omega_{\max}]$. It is remarkable in this context, that the weights q_1 and q_2 target at the respective forward whirl modes despite the system deviation, i.e. manipulation of these parameters has the desired effect on the respective resonance amplitudes. This statement cannot be considered proven by the results in this chapter, but is rather based on the experience the authors gained during controller design. However, good results achieved at both resonances using a single design point frequency being different from the respective resonance frequencies indicate the validity of this statement.

In this section, a proposal to reduce the space of free design parameters for the controller-observer combination to a manageable number of 5 was given, i.e. the parameters q_1, q_2, q_3, r_B and Ω_D are to be determined for controller design. However, and in contrast to [11], this number is too big for simple brute force optimization or manual tuning.

14.4 Controller Design via Genetic Optimization Algorithm

Following the proposal in [17], the 5 free design parameters are obtained using a GA [17–20] of the global optimization toolbox in MATLAB. Instead of Ω_D, $\tilde{\Omega}_D = \Omega_D/10$ is used as parameter of the optimization in order to achieve similar scaling of all free design parameters. In contrast to traditional gradient based optimization techniques, this stochastic global optimization algorithm can be applied to discontinuous objective functions and is intended to avoid getting stuck in local optima. Rather than working on a single point, GAs work on a whole population in parallel, starting with randomly distributed sets of solutions as the initial population. Every individual in a population possesses certain attributes i.e. the controller and observer design parameters in the case treated here, which are modified in order to minimize a certain objective function. Modification of these attributes is achieved by crossover and mutation in each iteration step based on a certain encoding of the attribute values, e.g. binary or real-valued encoding [20]. Mutation describes attribute alteration of randomly chosen individuals in isolation following certain rules, which of course differ for different encodings. Mutation guarantees that the probability of searching any region in the parameter space is never zero [19]. Crossover relies on the combination of attributes of certain pairs of individuals. The stochastic process of selection picks certain individuals from the population for crossover. The probability of a certain individual to get selected for crossover is determined on the basis of its objective function value. After the creation of new individuals by means of mutation and crossover, these are reinserted into the current population while keeping some elite, i.e. well-performing candidates from the previous population. In this way, a new generation is created in each iteration step.

A real-valued encoding is utilized here due to the facts that natural encoding of genetic information, which is to be emulated by the algorithm, is much more complicated than binary encoding and binary encoding furthermore possesses an undesirably high Hamming distance, which is disadvantageous for control of e.g. the mutation strength [20]. Hence, it may be assumed that the natural process of genetic manipulation is approximated in a more targeted manner using real-valued encoding.

The following objective function is used for optimization:

$$J_2 = \begin{cases} \infty, \text{closed loop is unstable for } \Omega^* \in [0;\ \Omega_{\max}] \\ \infty, \max(|u_i(\Omega^*)|) > u_{\max}, \Omega^* \in [0; \Omega_{\max}], i = 1, 2 \\ \infty, \left\| H_{u,\eta}(\Omega^*, s) \right\|_\infty > \gamma, \Omega^* \in [0; \Omega_{\max}] \\ \dfrac{1}{\Omega_{\max}} \displaystyle\int_0^{\Omega_{\max}} \sum_{i=1,2,\ldots,n} |x_i|^2 d\Omega, \quad \text{else} \end{cases} \tag{14.7}$$

$H_{(u,\eta)}(\Omega^*, s)$ denotes the FRF matrix from sensor noise inputs to actuator voltages of the system at a certain rotational frequency Ω^*. The value γ, as well as bounds for controller parameters are chosen on the basis of the controller in [11]. The authors recommend designing a controller via manual tuning of the 3 parameters of the controller structure proposed in [11] in advance to the optimization in order to get an idea of possible bounds when applying the strategy proposed here. In (14.7), the cost functional is set to ∞ if the system is unstable, actuator overload occurs or noise amplification is unacceptably high for any rotational frequency $\Omega^* \in [0; \Omega_{\max}]$ in the operating range. The choice of treatment of these cases via the objective function instead of constraints is due to the computationally involved routines required for treatment of nonlinear constraints in optimization problems, see e.g. [20]. Objective function values of ∞ can be handled by the algorithm due to the facts that discontinuities do not hamper genetic algorithms and ranking is utilized instead of e.g. proportional scaling functions in order to determine selection probability.

For the reasons already discussed in [11], the linear time variant stability proof is skipped. Instead, under the assumption of a slowly varying system, linear time invariant stability is checked for $\Omega^* \in [0; \Omega_{max}]$.

14.5 Results

The controller-observer combination is validated in simulation and experiment. Simulation results are achieved using steady state frequency domain analysis, while the measurement data was obtained during a slow run up at 50 rpm/s. In order to focus on the unbalance induced vibration, the measurement data is evaluated using a digital implementation of the wattmeter measuring principle [22] and extracting the first order vibration.

Vibration amplitudes at the two discs with and without control as well as the respective control effort are shown in Fig. 14.4. Vibration attenuation is excellent and considerably higher than in [11]. Simulation matches reality with high accuracy. However, the behavior around the second unbalance induced resonance is not fully replicated by the model, presumably due to progressive elastic characteristics at the rig, which are not included in the simulation. Furthermore, there is some residual mismatch in control effort prediction due to nonlinearity of piezoelectric actuators.

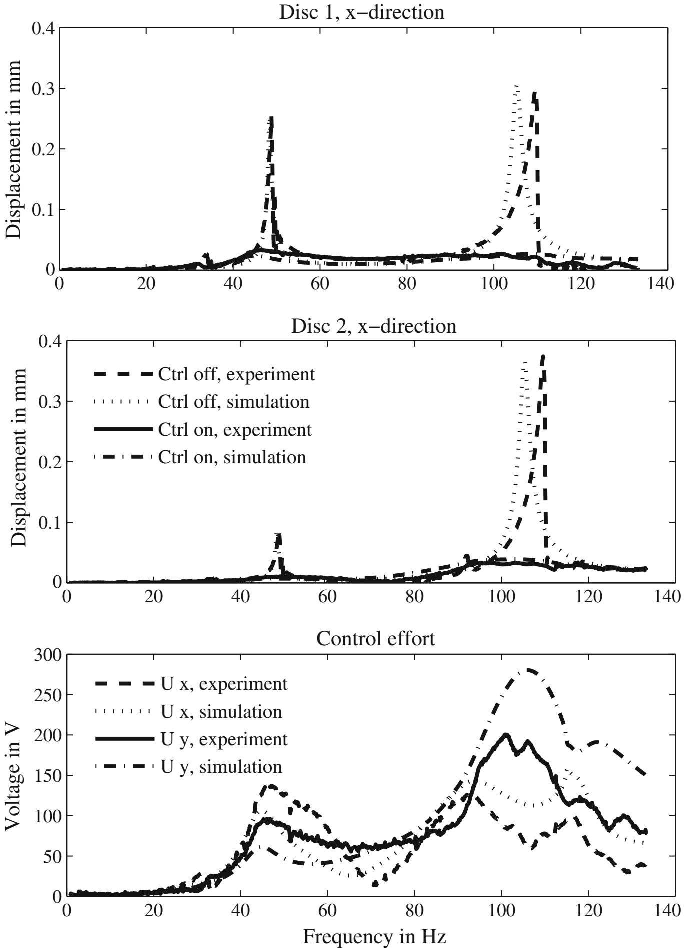


Fig. 14.4 Results of simulation and experiment, vibration amplitudes at the disc locations and control effort, data evaluated using the wattmeter measuring principle

14.6 Conclusion and Future Work

A proposal for combined design of a LQR and a Kalman Filter for a rotor with gyroscopic effect using a genetic optimization algorithm was given. The number of free design parameters for controller and observer design was reduced to a number of 5 on the basis of physical considerations. It was shown, that it is possible to design a single controller leading to robust stability and excellent vibration attenuation in the whole operating range, which covers two unbalance induced resonances. The controller was validated in simulation as well as experiment, and excellent results indicate, that the proposed controller design strategy is suitable for the problem. An easy and computationally efficient treatment of nonlinear constraints in the genetic algorithm could not be found in this investigation and is, besides other interesting control strategies, the topic of future research.

Acknowledgments This work is based on a research project in partnership with Rolls-Royce Deutschland Ltd & Co KG and was supported by Deutsche Forschungsgemeinschaft within the framework of the graduate college 1344, ‘‘Instationäre Systemmodellierung von Flugtriebwerken’’.

References

1. El-Shafei A (2002) Active control algorithms for the control of rotor vibrations using hybrid squeeze film dampers. J Eng Gas Turbines Power 124:598–607
2. Vance JM, Ying D, Nikolajsen JL (2000) Actively controlled bearing dampers for aircraft engine applications. J Eng Gas Turbines Power-Trans ASME 122(3):466–472
3. Tammi K (2009) Active control of rotor vibrations by two feedforward control algorithms. J Dyn Syst Meas Control 131:051012-1–051012-10
4. Fritto RL, Knospe CR (2002) Rotor compliance minimization via μ-control of active magnetic bearings. IEEE Trans Control Syst Technol 10(2):238–249
5. Balini HMNK, Scherer CW, Witte J (2011) Performance enhancement for AMB systems using unstable H∞ controllers. IEEE Trans Control Syst Tech 19(6):1–15
6. Tanaka N, Uchiyama N, Watanabe T, Seto K (2009) Levitation and vibration control of a flexible rotor by using active magnetic bearings. J Syst Des Dyn 3(4):551–562
7. Arias-Montiel M, Silva-Navarro G (2008) Finite element modeling and unbalance compensation for a two disks asymmetrical rotor system. In: Proceedings of the 5th international conference on electrical engineering, computing science and automatic control, Mexico City
8. Palazzolo AB, Lin RR, Alexander RM (1989) Piezoelectric pushers of active vibration control of rotating machinery. J Vib Acoust Stress Reliab Des 111:298–305
9. Lindenborn O, Hasch B, Nordmann R (2008) Vibration reduction and isolation of a rotor in an actively supported bearing using piezoelectric actuators and the FXLMS algorithm. In: Proceedings of the 9th international conference on vibrations in rotating machinery, Exeter
10. Lebo F, Rinderknecht S, Özel M (2012) Model-based control of an elastic aircraft engine rotor with piezo stack actuators. In: Proceedings of the IEEE 17th international conference on IE&EM, Xiamen
11. Schittenhelm RS, Borsdorf M, Riemann B, Rinderknecht S (2012) Linear quadratic regulation of a rotating shaft being subject to gyroscopic effect. In: Lecture notes in

engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS, San Francisco, USA, 24–26 Oct 2012, pp 1292–1297

12. Kuo SM, Morgan DR (1996) Active noise control systems. Wiley, New York
13. Preumont A (2004) Vibration control of active structures. Kluwer Academic Publishers, New York
14. Zhou K, Doyle JC, Glover K (1995) Robust and optimal control. Prentice Hall, New Jersey
15. Wang Z, Schittenhelm RS, Rinderknecht S (2012) Augmented observer for fault detection and isolation (FDI) in rotor systems. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS 2012, San Francisco, USA, 24–26 Oct 2012, pp 336–341
16. Hendricks E, Jannerup O, Sorensen PH (2008) Linear systems control. Springer, Berlin
17. Ejaz M (2012) Application of genetic algorithm for tuning of reduced ordered robust PID controller and computer simulations. Int J Comput Appl 42(17):20–24
18. El-Mihoub TA, Hopgood AA, Nolle L, Battersby A (2006) Hybrid genetic algorithms: a review. Eng Lett 13(2)
19. Chipperfield A, Fleming P, Pohlheim H, Fonseca C (1994) Genetic algorithm toolbox for use with MATLAB. Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield
20. Yu X, Gen M (2010) Introduction to evolutionary algorithms. Springer, London
21. Yamamoto T (2001) Linear and nonlinear rotordynamics. Wiley-Interscience, New York
22. Argeseanu A, Ritchie E, Leban K (2010) New balancing equipment for mass production of small and medium-sized electrical machines. In: Proceedings of the 12th international conference on optimization of electrical and electronic equipment, Brasov, pp 506–511

Chapter 15
Adaptive Control System for Solution of Fault Tolerance Problem

Yuri A. Vershinin

Abstract Adaptive control can provide desirable behavior of a process even though the process parameters are unknown or may vary with time. Conventional adaptive control requires that the speed of adaptation must be more rapid than that of the parameter changes. However, in practice, problems do arise when this is not the case. For example, when fault occurs in a process, the parameters may change very dramatically. A new approach based on simultaneous identification and adaptation of unknown parameters is suggested for compensation of rapidly changing parameters. High dynamic precision adaptive control can be used for the solution of a fault tolerance problem in complex and multivariable processes and systems.

Keywords Adaptive control · Fault tolerance · Hankel matrix · Identification · Markov parameters · Mathematical model · Singular value decomposition

15.1 Determination of a Mathematical Model of a Process

A mathematical model of a process on a stationary regime can be found from the sequence of Markov parameters using the classical Ho algorithm [1]. The Markov parameters can be obtained from input–output relationships or more directly as an impulse response of the system. It is well known that according to the theorem of Kronecker the rank of the Hankel matrix constructed from the Markov parameters is equal to the order of the system from which the parameters are obtained. Therefore, by consistently increasing the dimension of the Hankel matrix Γ until

Y. A. Vershinin (✉)
Department of Mechanical and Automotive Engineering, Coventry University,
Coventry, UK
e-mail: vershy@coventry.ac.uk

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_15,

$$\text{rank } \Gamma_r = \text{rank } \Gamma_{r+1}$$

the order of the system can be obtained as equal to r. However, in practical implementation, this rank-order relationship may not give accurate results due to several factors: sensitivity of the numerical rank calculation and bias of the rank if information about the process is corrupted by noise. This problem can be avoided using singular value decomposition (SVD) of the Hankel matrix:

$$\Gamma = USV^T, \tag{15.1}$$

where

$$U^T U = V^T V = I,$$

$$S = diag(\sigma_1, \sigma_2, \ldots, \sigma_l, \sigma_{l+1}, \ldots \sigma_n).$$

Here U and V are orthogonal matrices. The diagonal elements of the matrix S (the singular values) in (15.1) are arranged in the following order $\sigma_1 > \sigma_2 > \cdots > \sigma_n > 0$. Applying the property of SVD to reflect the order of a system through the smallest singular value, the order of the system can be determined with the tolerance required. From practical point of view a reduced order model is more preferable. Taking into account that the best approximation in the Hankel norm sense is within a distance of σ_{l+1}, the model of order l can be found. However, a relevant matrix built from Markov parameters of this reduced order model should also be of the Hankel matrix. But it is not an easy matter to find such a Hankel matrix for the reduced order process. A simpler solution, although theoretically not the best, can be found from the least squares approximation of the original Hankel matrix [2–4]. The discrete time state-space realization of the process can be determined from the relationship between Markov parameters and representation of the Hankel matrix through relevant controllability and observability matrices of the process:

$$\Gamma = \begin{bmatrix} C_d \\ C_d A_d \\ C_d A_d^2 \\ \cdot \\ \cdot \end{bmatrix} \begin{bmatrix} A_d & A_d B_d & A_d^2 B_d & \cdot & \cdot \end{bmatrix} = \Omega \mathrm{E}, \tag{15.2}$$

where

A_d is the system matrix,
B_d is the control matrix,
C_d is the output matrix,
Ω is the observability matrix,
E is the controllability matrix.

15.2 The Adaptive Control System

Consider a continuous time single input—single output second order plant (a process) given in the following canonical state space realization form:

$$\begin{aligned} \dot{x} &= A_c x + B_c u \\ y &= C_c x \end{aligned}, \tag{15.3}$$

where

$$A_c = \begin{bmatrix} 0 & 1 \\ a_{1p} & a_{2p} \end{bmatrix}, B_c = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, C_c = \begin{bmatrix} c_{1p} & c_{2p} \end{bmatrix},$$

u is the control signal,
y is the output of the plant.

Assume that at the time t parameters a_{1p} and a_{2p} change dramatically due to a fault in the system, but parameters c_{1p} and c_{2p} remain constant. The mathematical model of plant (15.3) can be represented in the following form:

$$\begin{aligned} \ddot{x}_p &= (\bar{a}_{2p} + \Delta a_{2p}(t))\dot{x}_p + (\bar{a}_{1p} + \Delta a_{1p}(t))x_p + u \\ y_p &= \bar{c}_{2p}\dot{x}_p + \bar{c}_{1p}x_p, \end{aligned}$$

where

$$\begin{aligned} a_{1p} &= \bar{a}_{1p} + \Delta a_{1p}(t), \\ a_{2p} &= \bar{a}_{2p} + \Delta a_{2p}(t), \end{aligned}$$

$\bar{a}_{1p}$, $\bar{a}_{2p}$, $\bar{c}_{1p}$, $\bar{c}_{2p}$ are the nominal parameters (constant) of the plant,
$\Delta a_{1p}(t)$, $\Delta a_{2p}(t)$ are the biases of the plant parameters (variable) from their nominal values,
x_p is the plant state,
y_p is the plant output.

A desirable behavior of the plant can be determined by the following reference model:

$$\begin{aligned} \ddot{x}_m &= a_{2m}\dot{x}_m + a_{1m}x_m + g \\ y_m &= c_{2m}\dot{x}_m + c_{1m}x_m \end{aligned}, \tag{15.4}$$

where

g is the input signal,
a_{1m}, a_{2m}, c_{1m}, c_{2m} are parameters of the model.

In order to compensate for the plant parameters' biases, a controller can be used. The closed loop system with the controller is represented in the following form:

$$\begin{aligned}\ddot{x}_p =&(\bar{a}_{2p} + \Delta a_{2p}(t))\dot{x}_p + (\bar{a}_{1p} + \Delta a_{1p}(t))x_p \\ &+ (\bar{k}_2 + \Delta k_2(t))\dot{x}_p + (\bar{k}_1 + \Delta k_1(t))x_p + g,\end{aligned} \tag{15.5}$$

where

$\bar{k}_1$, $\bar{k}_2$ are the constant parameters of the controller,
$\Delta k_1(t)$, $\Delta k_2(t)$ are the adjustable parameters of the controller.

The desirable quality of the process behavior can be obtained from the following relationships:

$$\begin{aligned}\bar{k}_1 + \bar{a}_{1p} &= a_{1m} \\ \bar{k}_2 + \bar{a}_{2p} &= a_{2m}.\end{aligned}$$

According to Eqs. (15.4) and (15.5), the error equation is obtained as follows:

$$\ddot{e} = a_{2m}\dot{e} + a_{1m}e + z_2\dot{x}_p + z_1x_p, \tag{15.6}$$

where

$$\begin{aligned}e &= x_m - x_p, \\ z_1 &= \Delta a_{1p}(t) + \Delta k_1(t), \\ z_2 &= \Delta a_{2p}(t) + \Delta k_2(t).\end{aligned}$$

It can be seen from Eq. (15.6) that in order to achieve the desirable error $e \to 0$, it is necessary to provide the following conditions:

$$z_1 \equiv 0,\ z_2 \equiv 0. \tag{15.7}$$

The conditions (15.7) can be achieved by adjusting parameters $\Delta k_1(t)$ and $\Delta k_2(t)$ according to the following laws [5]:

$$\begin{aligned}\Delta \dot{k}_1(t) &= \sigma x_p \\ \Delta \dot{k}_2(t) &= \sigma \dot{x}_p,\end{aligned} \tag{15.8}$$

where $\sigma = Pe$.

The positive definite symmetric matrix P can be obtained from the solution of the relevant Lyapunov equation. The main problem associated with algorithms (15.8) is that all self-tuning contours are linked through the dynamics of the plant. The consequence is that high interaction of each contour with others will occur. This further results in poor dynamic compensation of plant parameters' biases Δa_{ip} (i = 1, 2,…m), where m is a number of self-tuning contours. The idea of decoupling self-tuning contours from plant dynamics, based on simultaneous identification and adaptation, is suggested for the solution of this problem with fault tolerance. This could considerably improve performance of the overall system, especially for high dimension and multivariable plants and processes.

It can be shown [6, 7] that the self-tuning contours will be decoupled from the plant dynamics if σ can be formed such that:

$$\sigma^* = \ddot{e} - a_{2m}\dot{e} - a_{1m}e.$$

In this case the following relationship can be obtained:

$$\sigma^* = (\Delta a_{2p}(t) + \Delta k_2(t))\dot{x}_p + (\Delta a_{1p}(t) + \Delta k_1(t))x_p. \tag{15.9}$$

In order to solve Eq. (15.9) with two variable parameters, the following approach is suggested: Multiply both parts of Eq. (15.9) by state variables x_p and $\dot{x}_p$ and integrate the resultant equations on the time interval $(t_1,\ t_2)$, where: $t_2 = t_1 + \Delta t$. Taking the initial conditions as $t_1 = 0$, $\Delta k_i = 0$, $(i = 1, 2)$ the following equations are obtained:

$$\begin{aligned}\int_{t_1}^{t_1+\Delta t} \sigma^* x_p dt &= \Delta a_{2p} \int_{t_1}^{t_1+\Delta t} \dot{x}_p x_p dt + \Delta a_{1p} \int_{t_1}^{t_1+\Delta t} x_p^2 dt\\ \int_{t_1}^{t_1+\Delta t} \sigma^* \dot{x}_p dt &= \Delta a_{2p} \int_{t_1}^{t_1+\Delta t} \dot{x}_p^2 dt + \Delta a_{1p} \int_{t_1}^{t_1+\Delta t} x_p \dot{x}_p dt.\end{aligned} \tag{15.10}$$

Introduce the following notations:

$$\begin{gathered}\int_{t_1}^{t_1+\Delta t} \sigma^* x_p dt = c_1, \int_{t_1}^{t_1+\Delta t} \sigma^* \dot{x}_p dt = c_2,\\ \int_{t_1}^{t_1+\Delta t} x_p^2 dt = l_{11}, \int_{t_1}^{t_1+\Delta t} x_p \dot{x}_p dt = l_{21},\\ \int_{t_1}^{t_1+\Delta t} \dot{x}_p x_p dt = l_{12}, \int_{t_1}^{t_1+\Delta t} \dot{x}_p^2 dt = l_{22}.\end{gathered} \tag{15.11}$$

According to notations (15.11), Eq. (15.10) can now be written in the form:

$$\begin{aligned} c_1 &= \Delta a_{1p} l_{11} + \Delta a_{2p} l_{12}\\ c_2 &= \Delta a_{1p} l_{21} + \Delta a_{2p} l_{22}.\end{aligned} \tag{15.12}$$

From the solution of Eq. (15.12) the bias of the plant parameters Δa_{ip}, $(i = 1, 2)$ can be determined. The controller can be adjusted according to the estimated parameter bias as:

$$\Delta k_i = -\Delta a_{ip}.$$

Therefore, conditions (15.7) are satisfied, which in turn means that the behavior of system (15.5) follows the desirable trajectories of model (15.4), even in the presence of dramatic plant parameters changes.

For the solution of Eq. (15.12) one needs to take into account of the hypothesis of quasi-stationarity of the process, where the interval time Δt is selected such that the biases of parameters Δa_{ip} must be constant at this interval. However, the interval Δt should be sufficiently large in order to accumulate a larger quantity of variables x_p and $\dot{x}_p$ for the solution of the equations.

15.3 The Numerical Results

The Hankel matrix Γ, constructed from the Markov parameters (obtained from the experiment, see Appendix), is as follows:

$$\Gamma = \begin{bmatrix} 6.5000000\mathrm{e}-02 & 1.4550000\mathrm{e}-01 & 1.6442500\mathrm{e}-01 \\ 1.4550000\mathrm{e}-01 & 1.6442500\mathrm{e}-01 & 1.5056000\mathrm{e}-01 \\ 1.6442500\mathrm{e}-01 & 1.5056000\mathrm{e}-01 & 1.2447038\mathrm{e}-01 \end{bmatrix}. \tag{15.13}$$

Applying the singular value decomposition procedure (15.1) on the Hankel matrix (15.13), it is found that

$$U = \begin{bmatrix} 5.1633320\mathrm{e}-01 & 8.1190203\mathrm{e}-01 & 2.7242453\mathrm{e}-01 \\ 6.2194166\mathrm{e}-01 & -1.3682059\mathrm{e}-01 & -7.7101797\mathrm{e}-01 \\ 5.8871776\mathrm{e}-01 & -5.6753434\mathrm{e}-01 & 5.7560070\mathrm{e}-01 \end{bmatrix}$$

$$V = \begin{bmatrix} 5.1633320\mathrm{e}-01 & -8.1190203\mathrm{e}-01 & 2.7242453\mathrm{e}-01 \\ 6.2194166\mathrm{e}-01 & 1.3682059\mathrm{e}-01 & -7.7101797\mathrm{e}-01 \\ 5.8871776\mathrm{e}-01 & 5.6753434\mathrm{e}-01 & 5.7560070\mathrm{e}-01 \end{bmatrix}$$

$$S = \begin{bmatrix} 4.2773559\mathrm{e}-01 & 0.0000000\mathrm{e}+00 & 0.0000000\mathrm{e}+00 \\ 0.0000000\mathrm{e}+00 & 7.4455532\mathrm{e}-02 & 0.0000000\mathrm{e}+00 \\ 0.0000000\mathrm{e}+00 & 0.0000000\mathrm{e}+00 & 6.1531296\mathrm{e}-04 \end{bmatrix}. \tag{15.14}$$

Using relations (15.1), (15.2) and (15.14) the discrete time state space realization of the reduced order system is obtained as follows:

$$\begin{gathered} A_d = \begin{bmatrix} 9.7950468\mathrm{e}-01 & -3.4211654\mathrm{e}-01 \\ 3.4211654\mathrm{e}-01 & 3.4867831\mathrm{e}-01 \end{bmatrix} \\ B_d = \begin{bmatrix} 3.3767560\mathrm{e}-01 \\ -2.2160613\mathrm{e}-01 \end{bmatrix} \\ C_d = \begin{bmatrix} 3.3767560\mathrm{e}-01 & 2.2160613\mathrm{e}-01 \end{bmatrix} \end{gathered} \tag{15.15}$$

The behavior of the full order model and the reduced order model is given in Fig. 15.1. It can be seen in Fig. 15.1 and Appendix that the Markov parameters of

Fig. 15.1 The behavior of the full order model and reduced order model

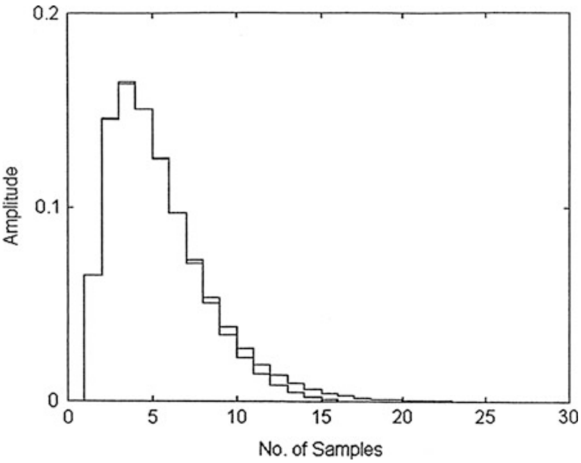
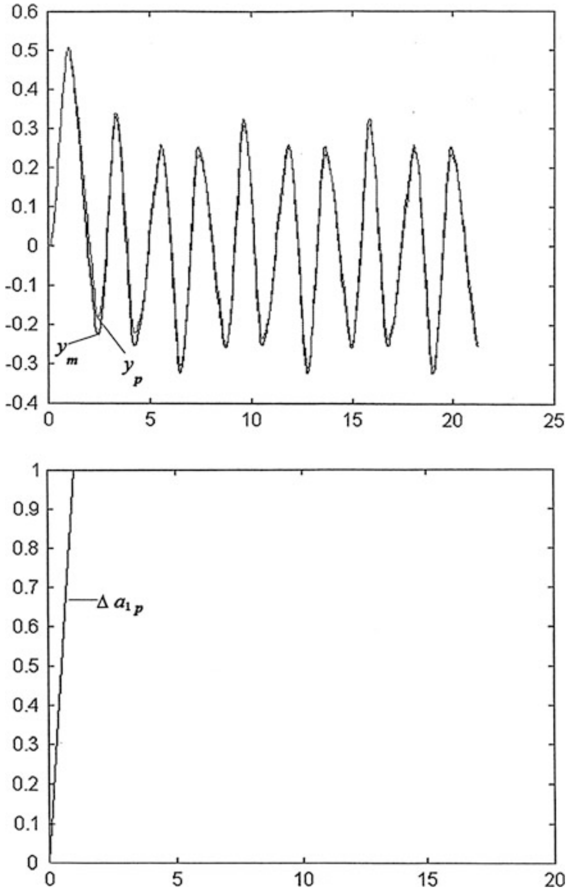


Fig. 15.2 Bias $\Delta a_{1p} = 1$, $\Delta a_{2p} = 0$. The adaptation is switched off



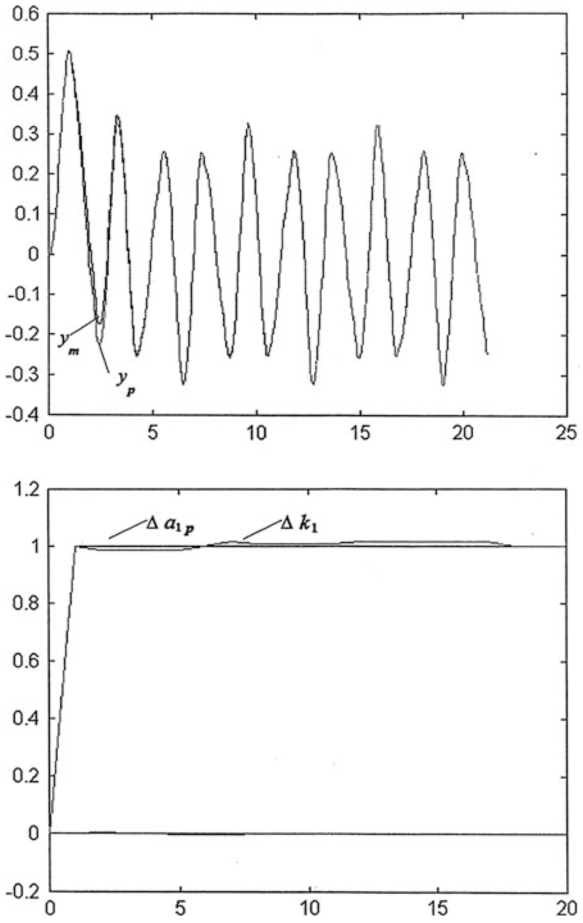


Fig. 15.3 Bias $\Delta a_{1p} = 1$, $\Delta a_{2p} = 0$. The adaptation is switched on

the reduced order model are a close approximation to the Markov parameters of the original system.

Nominal parameters of the plant in the continuous time (15.3) are obtained from (15.15) as follows:

$$\bar{a}_{1p} = -3.1184,\ \bar{a}_{2p} = -3.0517,$$
$$\bar{c}_{1p} = -0.0318,\ \bar{c}_{2p} = 2.9132.$$

Parameters of model (15.4) are chosen as $a_{1m} = \bar{a}_{1p}$, $a_{2m} = \bar{a}_{2p}$, $c_{1m} = \bar{c}_{1p}$, $c_{2m} = \bar{c}_{2p}$.

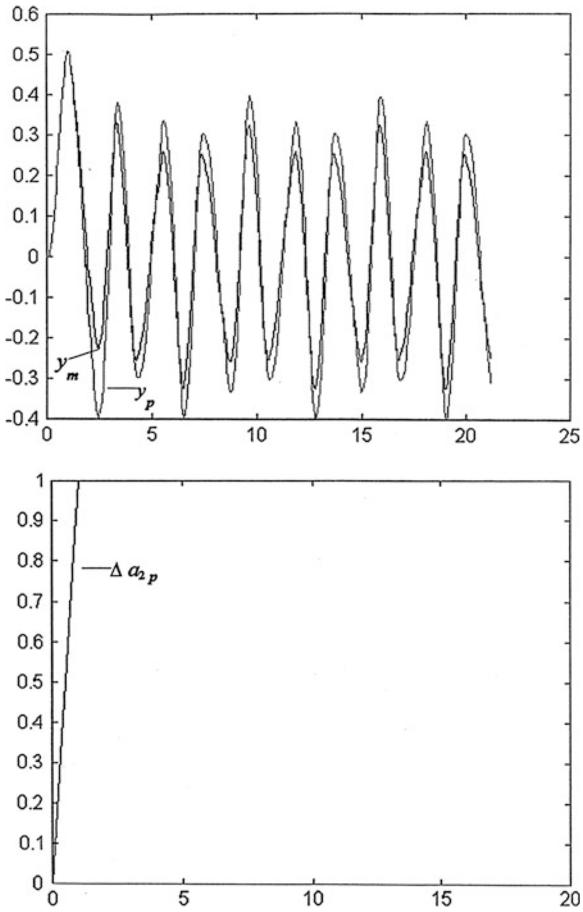


Fig. 15.4 Bias $\Delta a_{1p} = 0$, $\Delta a_{2p} = 1$. The adaptation is switched off

The performance of the high dynamic precision adaptive control system is presented in Figs. 15.2, 15.3, 15.4, 15.5.

Figure 15.2 shows that the bias from the nominal parameter at time $t \geq 1$ s is $\Delta a_{1p} = 1$, ($\Delta a_{2p} = 0$). The adaptation is switched off.

Figure 15.3 shows the bias from the nominal parameter at $t \geq 1$ s with adaptation being switched on ($\Delta a_{1p} = 1$, $\Delta a_{2p} = 0$). It can be seen that the output of system y_p coincides with the model reference output y_m after $t \geq 4$ s.

Figure 15.4 shows that the bias from the nominal parameter at time $t \geq 1$ s is $\Delta a_{2p} = 1$, ($\Delta a_{1p} = 0$). The adaptation is switched off.

Figure 15.5 shows the bias from the nominal parameter at $t \geq 1$ s with adaptation being switched on ($\Delta a_{2p} = 1$, $\Delta a_{1p} = 0$). It can be seen that the output of system y_p coincides with the model reference output y_m after $t \geq 9$ s.

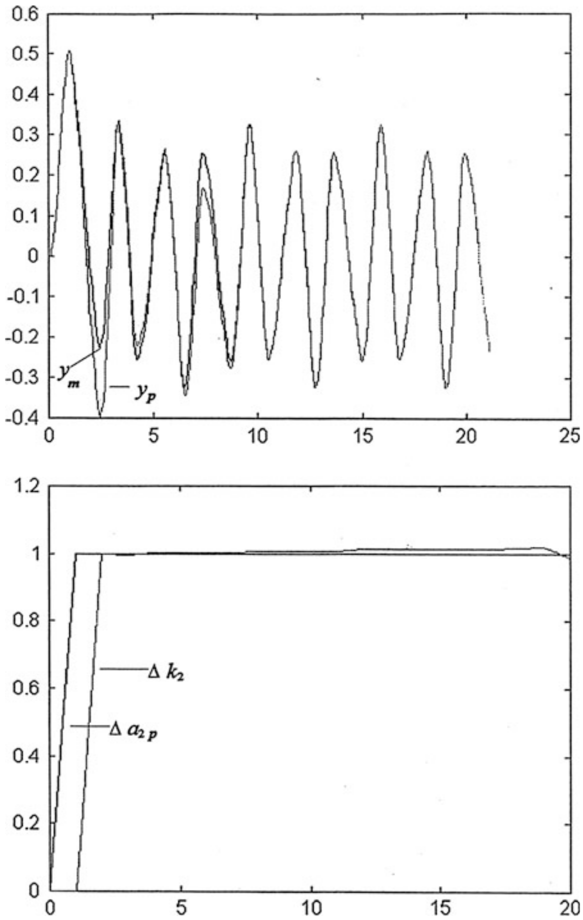


Fig. 15.5 Bias $\Delta a_{1p} = 0$, $\Delta a_{2p} = 1$. The adaptation is switched on

15.4 Conclusions

The high dynamic precision adaptive control system for the solution of a fault tolerance problem of a single–input–single–output process is suggested in this paper. The method, which is based on simultaneous identification and adaptation of unknown process parameters, provides decoupling of self-tuning contours from plant dynamics. The control system compensates the rapidly changing parameter when fault occurs in a process. The mathematical model of the process is formed from Markov parameters, which are obtained from the experiment as the process impulse response. The order of the model is determined using singular value decomposition of the relevant Hankel matrix. This allows one to obtain a robust reduced order model representation if the information about the process is corrupted by noise in industrial environment. The adaptive control can be used for the solution of a fault tolerance problem [8] in complex and multivariable processes and systems.

Appendix

Markov parameters obtained from the experiment	Markov parameters of the reduced order model
0.0000000e + 00	0.0000000e + 00
6.5000000e − 02	6.4934730e − 02
1.4550000e − 01	1.4578163e − 01
1.6442500e − 01	1.6384913e − 01
1.5056000e − 01	1.5077128e − 01
1.2447038e − 01	1.2511681e − 01
9.7003263e − 02	9.7037520e − 02
7.2809279e − 02	7.1509116e − 02
5.3273657e − 02	5.0478548e − 02
2.7143404e − 02	3.4252666e − 02
1.9054881e − 02	2.2345734e − 02
1.3274250e − 02	1.3971877e − 02
9.1920232e − 03	8.3100499e − 03
6.3351771e − 03	4.6301281e − 03
4.3498142e − 03	2.3388797e − 03
2.9776238e − 03	9.8319708e − 04
2.0333343e − 03	2.3330942e − 04
1.3857582e − 03	−1.4099694e − 04
9.4289895e − 04	−2.9426412e − 04
6.4072233e − 04	−3.2618265e − 04

References

1. Ho, BL, Kalman RE (1966) Effective construction of linear state-variable models from input/output functions. In: Proceedings the third Allerton conference, pp 449–459
2. Zeiger HP, McEwen AJ (1974) Approximate linear realizations of given dimension via Ho's algorithm. IEEE Trans Autom Control 19:153
3. Kalman RE, Falb PL, Arbib MA (1974) Topics in mathematical system theory, McGraw Hill, New York
4. Moor BC (1981) Principal component analysis in linear systems: Controllability, observability and model reduction. IEEE Trans Autom Control 26:17–32
5. Astrom KJ, Wittenmark B (1995) Adaptive control. Addison Wesley, Reading, Mass, Boston
6. Petrov BN, Rutkovsky VY, Zemlyakov SD (1980) Adaptive coordinate-parametric control of non-stationary plants. Nauka, Moscow
7. Vershinin YA (1991) Guarantee of tuning independence in multivariable systems. In: Proceedings of the 5th Leningrad conference on theory of adaptive control systems: adaptive and expert systems in control, Leningrad
8. Vershinin YA (2012) Application of self-tuning control system for solution of fault tolerance problem, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, USA, 24–26 Oct 2012, pp 1206–1210

Chapter 16
Multiple Direction-of-Arrival Estimation for a Mobile Robotic Platform with Small Hardware Setup

Caleb Rascon and Luis Pineda

Abstract Knowledge of how many users are there in the environment, and where they are located is essential for natural and efficient Human-Robot Interaction (HRI). However, carrying out the estimation of multiple Directions-of-Arrival (multi-DOA) on a mobile robotic platform involves a greater challenge as the mobility of the service robot needs to be considered when proposing a solution. This needs to strike a balance with the performance of the DOA estimation, specifically the amount of users the system can detect, which is usually limited by the amount of microphones used. In this contribution, an appropriately carriable small and lightweight hardware system (based on a 3-microphone triangular system) is used, and a fast multi-DOA estimator is proposed that is able to estimate more users than the number of microphones employed.

Keywords HRI · Lightweight · Microphone array · Mobile · Multiple direction of arrival · Reverberation · Service robot

16.1 Introduction

The problem known as Multiple Direction-of-Arrival (multi-DOA) Estimation provides a unique challenge when being carried out in a mobile hardware platform, such as service robots. However, it plays an essential part of a natural Human-Robot Interaction (HRI), as it is important to know from where the users are talking to the robot and how many are there in the environment.

C. Rascon (✉) · L. Pineda
Universidad Nacional Autónoma de México, Mexico City, México
e-mail: caleb.rascon@iimas.unam.mx

L. Pineda
e-mail: lpineda@unam.mx

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_16,

From the technical point of view, knowing the direction of the user in relation to the robot can benefit other system modules. For instance, once the direction of the user is known, voice recognition can be improved using directional noise cancellation [1] or by simply turning a directional microphone in the direction of the user.

In addition, it is well known that face detection and recognition provide rich information relevant to HRI: the identity of the user, the direction the user is looking at, his mood, etc. [2, 3]. However, such analysis is carried out by visual means, and the user cannot always be expected to be in the line of sight of the robot. When dealing with human beings, by sorts of providence, the mouth is expected to be in the vicinity of the face of the user, which means that knowing the direction of the user by speech alone also provides a good heuristic of the location of his/her face. Using this information to face the user tackles the visual-range-limitation issue straight on.

Moreover, the robot may be expected to be in a situation where several users are in the environment and actively speaking to the robot, such as taking a food order or while guiding a group of users in a tour. Knowledge of the amount of users and from where are they talking to the robot can be used to provide acoustic cues to separate several streams of audio data from the environment based on the Direction-of-Arrival (DOA) of the various sound sources and provide the single-source streams to the Automatic Speech Recognizer. This provides the functionality of being able to carry out ASR of multiple users that may be interrupting each other, an occurrence bound to happen in a multiple-user scenario.

From the perspective of the user, the action of the robot facing him/her when being talked to acts as a type of bodily feedback which the user will naturally interpret as if the robot is ‘putting attention’ to him/her. This interpretation is an important part of a successful HRI, as the robot reacts in a way expected by the user and, at the same time, provides important feedback that makes the user feel acknowledged at the very beginning of the interaction. Meaning that, with only this seemingly trivial act, a good preamble to HRI is put forward.

In addition, the location of the user is an important variable in HRI. During a human-robot conversation, the phrase “robot, come here” may be emitted by the human. In this situation, even if the phrase was recognized correctly, the robot may know that it needs to move, but, because the term ‘here’ lacks context, it will not know **where** to move. Knowing the direction of the user in regard to the robot is an essential variable in the estimation of the location of the user in the environment. In a 3-dimensional polar coordinate system, the horizontal angle (i.e. the direction of the user) is one of three values that define a location (the other two being: vertical angle and distance from origin). Using heuristics from the environment, the DOA of the user can be used to segregate the locations where the user is most probably at. This means that when using ASR and DOA estimation conjunctively, the aforementioned phrase can be contextualized and stripped of its vagueness. From the user’s point of view, only a vague command is enunciated and the robot is able to carry it out, which is more ‘natural’ for the user than to position themselves in front of the robot.

Unfortunately, there are many challenges in the estimation of the DOA of the sound source. Reverberation is prevalent in the locations where a service robot is expected to be (supermarkets, restaurants, condominiums, etc.) and has been shown to hinder considerably the effectiveness of current DOA estimators [4]. Moreover, too many sound sources may drown the acoustic environment, complicating the estimation process. A sophisticated audio capturing system may be able to overcome these issues, such as the one proposed in [5] that used a 24-microphone 1-D array for precision. However, the application landscape of service robots provides a unique challenge for the multi-DOA estimation topic: a high amount of microphones may be impractical to carry by many of the currently-in-use service robots [6, 7], such as our in-house robot, Golem-II+, herein described.

Golem-II+ is a service robot built with a primary focus on HRI. It is integrated by a cognitive architecture focused on HRI, termed Interaction-Oriented Cognitive Architecture (IOCA) [8], which can take advantage of different types of information interpreted from the world, including the direction of the user. Because Golem-II+ is a conversational robot, it is of interest that it is able to detect and carry out conversations with several users at any point. This implies that the system that is to be estimating the multiple DOAs of the environment, needs to be sufficiently light on the hardware side for the robot to carry and not hinder its mobility, but robust enough in the software side to handle different types of noise and disturbances, as well as simultaneous speech from various sources. Moreover, such a system should be able to estimate the direction of the users in a $-179°$–$180°$ range, as no assumption can be made of the location of the users in the environment, and fast enough to do so with small utterances from the users. It is important to note, then, that the Multi-DOA Estimation problem is further complicated in a mobile robotic platform, and provides an interesting and unique challenge for current techniques.

This contribution is organized as follows: Sect. 16.2 is a brief review of current algorithms that aim to estimate the direction of one or more sound sources; Sect. 16.3 describes the proposed system; in Sect. 16.4, the results of the evaluation of the system on a service robot placed on a highly acoustically-complex scenario are provided; and in Sect. 16.5, conclusions and future work are discussed.

16.2 Background on Source Direction-of-Arrival Estimation

Estimating a Sound Source Direction of Arrival (DOA) is a well written-about topic in Signal Processing. It has been proven useful in applications ranging from fault monitoring in aircrafts [5], to intricate robotic pets [9], to close-to-life insect emulation [10]. In addition, the principles employed in DOA estimation have been applied in the design of hearing aids [1].

Having two audio sensors (i.e. microphones), the Inter-aural Time Difference (ITD) is the delay of a sound from one microphone to the other. Its calculation is usually based on the Cross-Correlation Vector (CCV) between the two captured signals. One of the simplest way to calculate the CCV is by applying Eq. (16.1).

$$CCV[k] = \frac{\sum_i (x_i - m_x)(y_{i-k} - m_y)}{\sqrt{\sum_i (x_i - m_x)^2}\sqrt{\sum_i (y_{i-k} - m_y)^2}} \tag{16.1}$$

where x and y are the two discrete signals being compared; k is the point at which y is being linearly shifted and the correlation is being calculated; m_x and m_y are the mean values of x and y, respectively. The ITD is the k value of the highest correlation measure in the CCV. It is one the features most used for DOA estimation, particularly with two-microphone arrays, as in [11] where it provided limited results. The ITD yields a clear relation to the direction of the source, described in Eq. (16.2).

$$\theta = arcsin\left(\frac{ITD \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{16.2}$$

where θ is the DOA angle; ITD is the Inter-aural Time Difference in number of samples; V_{sound} is the speed of sound ($\sim 343\,\mathrm{m/s}$); F_{sample} is the sampling frequency; and d is the distance between microphones.

The Inter-aural Intensity Difference (IID) is the difference in magnitude between both microphones and can also be used for DOA estimation, although a training stage is usually necessary for it to be useful, as it was observed in [7].

In [4], the concept of Inter-aural Coherence (IC) is introduced, which is the highest correlation measure of the CCV. If a high IC is present, the signals are deemed coherent and, thus, an analysis using ITD and/or IID can proceed. This methodology was implemented in [9], and it was observed that it didn't improve DOA estimation when dealing with complex signals (e.g. more than one source, reverberation present, etc.).

A popular methodology for DOA estimation in robotic platforms is to use a microphone array with, usually, two microphones, as it is proposed in [12]. The reasoning behind using only two microphones in robotic platforms ranges from that of practicality (it is lightweight), to that of biological similarity [13, 14] where the robot is meant to be the most human-like possible. However, doing so comes with four main problems.

ITD-DOA Non-Linear Relation. In Fig. 16.1, the DOA is plotted against the ITD, and it can be seen that in the $-50°$–$50°$ range, the relation between both seem close-to-linear. However, in the outer ranges, the relation becomes exponential. This causes major errors when estimating angles that are located in the sides of the robot [12]. This issue can be overcome by only estimating DOAs in the linear range, but, as it will be described, the DOA range is already limited as it is.

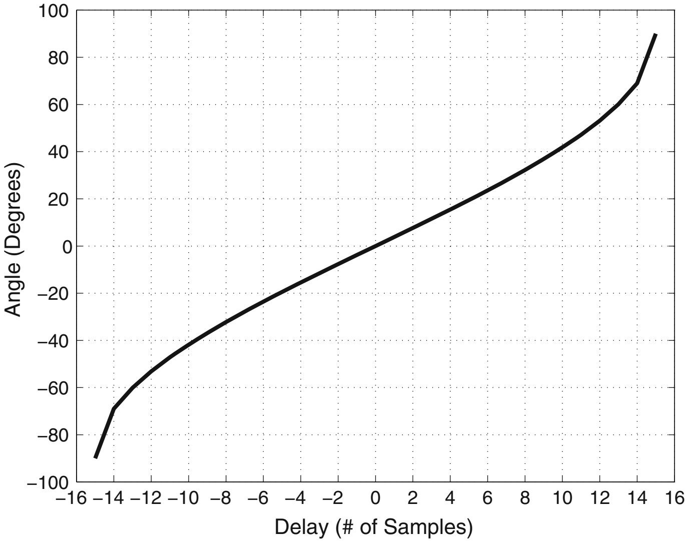


Fig. 16.1 DOA (or *Angle*) in degrees versus ITD (or *Delay*) in # of samples

Limited DOA Range. As it can also be seen in Fig. 16.1, a 2-mic array only estimates DOAs in the −90°–90° range. This can be surmounted by implementing 'artificial ears' that can detect if the sound source is coming from the front or back of the robot, but it has been proven impractical [15]. This can also be tackled by a two-phase strategy: a first pair of signals can be used to estimate an initial DOA, the robot can then rotate briefly, and then another pair of signals can be acquired to estimate a second DOA. A comparison between the DOAs results in an angle estimation in the −179°–180° range, but has its own set of issues: it requires considerably more time than when using one DOA estimate, the required rotation may hinder navigational requirements, and the user may be moving as well, rendering the DOA comparison mute.

Reverberation Sensitivity. The estimation of the ITD, based on the calculation of a CCV, can be very sensitive to reverberations and other noise sources [13, pp. 213–215]. This may result in significant errors in the DOA estimation without any form of redundancy.

Number of Microphones. A 2-microphone array has rarely been used for multi-DOA estimation, as it provides sparse information from the environment. Adding more microphones generalizes the strategy, as a 2-microphone array is an instantiation of classic reverse beamforming techniques [5], which create a noise map of the environment, and then, by using metrics such as energy levels, propose possible sources of sound and their respective DOAs. However, to obtain a high resolution noise map, and, thus, a precise DOA estimation, beamforming techniques require a large quantity of microphones, which is impractical for mobile robotic platforms. In addition, the more popular 1-dimensional (1-D) beamforming methodologies are also bounded by the first three problems described earlier, and 2-D arrays can be cumbersome to the mobility of the robot.

The topic of how many microphones to use in a service robot is intrinsic to the nature of the application, as it is important for the audio capture system to be mobile. A many-microphone solution may provide good results, such as the one proposed in [6] where the sources were separated from each other, in order to enhance speech recognition, and as a preamble for DOA estimation. However, it required an array of 8 microphones positioned in a cube-like manner to work, doubling the height the robot occupied without it.

The other side of the argument is to use one microphone, such as the work described in [15], where the DOA of a source was able to be estimated by implementing an 'artificial ear'. Unfortunately, the sound was required to be known *a priori* and any modification to the ear (even its location in relation to the microphone) required re-training.

A popular technique is the Multiple Signal Classification algorithm (MUSIC) [16], which is able to detect the Direction of Arrival (DOA) of as many sources as one less the available microphones (e.g. 1 source with 2 microphones, 2 sources with 3 microphones, etc). It does this by projecting the received signals in a DOA subspace, based on their eigenvectors, similar to Principal Component Analysis. It was applied in [17] with good results, although it has been observed that its performance decreases considerably in the presence of reverberation [13, pp. 169].

In this contribution, a technique is proposed where a small hardware system (based on only 3 microphones) is able to estimate multiple DOAs, as much as 4 simultaneous sources.

16.3 Proposed System

The work carried out in [18], which, in turn, is based on the proposal presented in [19] is the basis of the proposed system in this contribution. It is comprised by three modules that are described extensively in the rest of this section:

1. *Audio Acquisition*. Obtains audio data from the microphones and provides it to the Initial DOA Estimation module.
2. *Initial DOA Estimation*. Estimates, from the audio data, an initial, fast, but reliable DOA estimation of a single sound source in the environment.
3. *Multi-DOA Tracking*. Carries out dynamic clustering of the incoming DOA estimations, and proposes clusters of DOAs as sound sources.

16.3.1 Audio Acquisition

As it will be described in the following section, the hardware is comprised by three omnidirectional microphones, and, because the DOA estimation is based on an

ITD measure, it requires that the audio from the three microphones be acquired simultaneously as well as in real-time. For this purpose, the JACK Audio Connection Toolkit [20] was employed, as it can sample at rates of 44.1 kHz and 48 kHz, providing a good resolution for ITD calculations, and it does so without slowing down the other robotic software modules.

16.3.2 Initial DOA Estimation

The Initial DOA Estimation is carried out by the technique described in [19]. It avoids the problems that arise when estimating a DOA using 1-D microphones arrays (described in Sect. 16.2), and maintains a relatively small hardware setup: an equilaterial-triangular-array, as it is shown in Fig. 16.2. To this effect, the system obtains a set of 3 simultaneous sample windows.

The audio data is passed through various serialized sub-modules: a band-pass filter, a Voice Activity Detection stage, multi-ITD estimation, a redundancy check, and, finally, a final DOA estimation. The flow of data is summarized in Fig. 16.3.

16.3.2.1 Band-Pass Filter

A general infinite impulse response band-pass filter is used at the beginning of the process, to remove general ambient noise that is outside the human speech frequency bands. The filter model is described in Eq. (16.3):

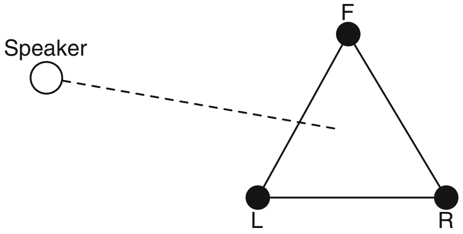


Fig. 16.2 Hardware setup of the proposed system

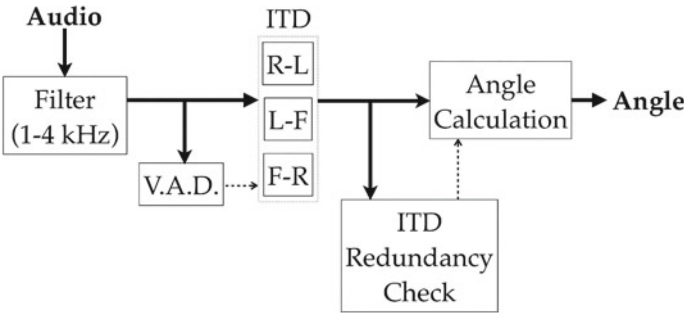


Fig. 16.3 Initial DOA Estimation flow of data

$$\begin{aligned} y_n = {} & 0.0348 \cdot x_n - 0.0696 \cdot x_{n-2} + 0.0348 \cdot x_{n-4} \\ & + 3.2680 \cdot y_{n-1} - 4.1247 \cdot y_{n-2} + 2.3984 \cdot y_{n-3} - 0.5466 \cdot y_{n-4} \end{aligned} \tag{16.3}$$

where y_i is the output of the filter, x_i is the input, and n is the number of the current sample.

It was observed that this filter made the system less sensitive towards unwanted noises that should always be ignored, such as high-pitch sounds, microphone hiss, etc. Concurrently, it did not degrate the sensitivity of the system towards human speech.

16.3.2.2 Voice Activity Detection

To trigger the ITD estimation described in the next section, Voice Activity Detection (VAD) needs to be carried out. Because the robotic platform may be changing position and environments, the VAD system is required to adjust to such changes automatically. To this effect, a simple VAD algorithm is proposed that is based on adjusting the baseline of the environmental noise to any sound that is emitted with a pre-specified delay.

Two history buffers of sample window energy values are kept in memory and shifted based on the specified delay (2 s provided good results). One buffer is always being refreshed by new sample window energy values (*avg_buffer*), and is used to calculate the current average energy value (*avg_value*). The other buffer (*min_buffer*) is used to calculate the current average minimum value (*min_value*), and is refreshed with a new energy value if it is less than the current *min_value* or if the difference between it and *avg_value* is less than the difference between *avg_value* and *min_value* (which would mean that its value is close to the values of *min_buffer*).

The VAD is triggered if the energy value of the current sample window is greater than the average between *avg_value* and *min_value* by a multiplicative threshold (1.5 provided good results).

16.3.2.3 Multi-ITD Estimation, Redundancy Check, and Angle Calculation

Once the VAD is triggered, the Multi-ITD Estimation follows. Three possible ITDs can be calculated using cross-correlation between sample window R and L (I_{RL}), L and F (I_{LF}), and F and R (I_{FR}). 2 DOAs are calculated from each ITD: one using Eq. (16.2), and another shifting the first DOA to its possible counterpart on the 'backside' of the microphone pair.

The three DOA pairs are used to check if the three ITDs are from a sound source located in the same angle sector. To do this, the average of the differences between the DOA pairs is calculated using Eq. (16.4).

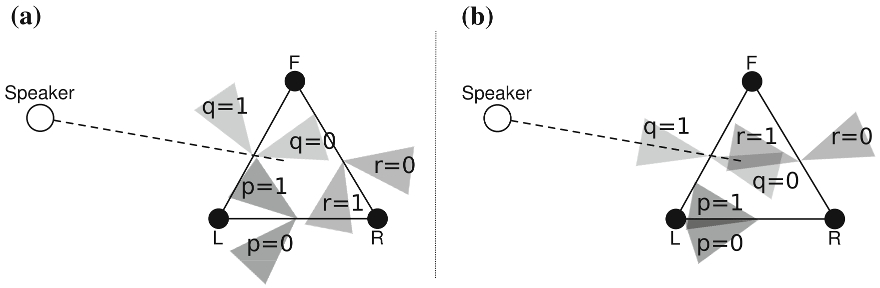


Fig. 16.4 **a** A highly incoherent ITD set. **b** A coherent ITD set ($p = q = r = 1$)

$$C_{pqr} = \frac{|D^p_{RL} - D^q_{LF}| + |D^q_{LF} - D^r_{FR}| + |D^r_{FR} - D^p_{RL}|}{3} \tag{16.4}$$

where a D^i_{xy} is the ith DOA of the DOA pair from I_{xy}. A set of 8 C_{pqr} are calculated, where p, q, and r can be either 0 or 1, depending on which DOA of the DOA pair is being compared. Of the 8, the minimum is considered as the *incoherence* of the sample window set. As it can be seen in Fig. 16.4a there is no combination of p, q, and r DOAs that provide low incoherence, while in Fig. 16.4b, the combination $p = q = r = 1$ provides good coherence, all three pointing towards the source.

A pre-specified *incoherence threshold* (measured in degrees of separation between the DOAs; a value between 30° and 40° provided good results) is used to reject sample window sets. A high incoherence implies that the sample window set either has too much reverberation to be trustworthy for further processing or that it contains **more than one sound source**. This rejection step serves as a type of redundancy check *per sampling window set.*

If all of the DOAs are coherent/redundant with each other, a preliminary DOA value (θ_m) can be calculated using Eq. (16.5),

$$\theta_m = arcsin\left(\frac{I_{min} \cdot V_{sound}}{F_{sample} \cdot d}\right) \tag{16.5}$$

where I_{min} is the ITD with the lowest absolute value of the three (I_{RL}, I_{LF}, I_{FR}). θ_m is then shifted to the appropriate angle sector in relation to the orientation of the robot, resulting in the final DOA value (θ).

Using I_{min} ensures that θ_m is calculated from the microphone pair that is the most perpendicular to the source. This means that the resulting θ is always estimated using a θ_m inside the −30°–30° range (well within the close-to-linear −50°–50° range), because of the equilateral nature of the triangular array. Meaning that all through the −179°–180° range, there is always a close-to-linear ITD-DOA relation.

Because of both the redundancy check and the close-to-linear relation, the maximum error of this system can be known beforehand using Eq. (16.6).

$$|error^{\circ}_{max}| = arcsin\left(\frac{I_{>30^{\circ}} \cdot V_{sound}}{F_{sample} \cdot d}\right) - arcsin\left(\frac{I_{<30^{\circ}} \cdot V_{sound}}{F_{sample} \cdot d}\right) \qquad (16.6)$$

where $I_{>30^{\circ}}$ and $I_{<30^{\circ}}$ are the ITDs that provide the closest ceil and floor measurements, respectively, to 30°. For example, sampling at 44.1 kHz and with the microphones spaced at 18 cm, a maximum error of ±2.8747° can be expected. In the same set of circumstances, when using a 2-Mic 1-D array, the maximum expected error, which occurs when the sound source is close to either side of the robot, is of ±15.0548°.

16.3.3 Multi-DOA Tracking

The DOA estimator described in the previous section only provides results when there is considerable confidence of only one sound source being detected in a small sample window (up to 100 ms). Although, it has been shown that people tend to not talk over each other while in conversation [21], even in simultaneous-speech, it has been seen that users are not expected to talk with a 100 % overlap over each other. In fact, when analyzing speech recognition, 'spurts' of non-overlapping speech has been considered to the order of 500 ms [22]. For example, in Fig. 16.5, it can be seen how two randomly chosen tracks from the DIMEX corpus [23], when overlayed over each other, still have some portions with no overlap between them.

This means that the DOA estimator described in the last section is able to provide reliable results of single sources even in multi-user scenarios. However, because of the stochastic nature of the presence of single user sample windows in

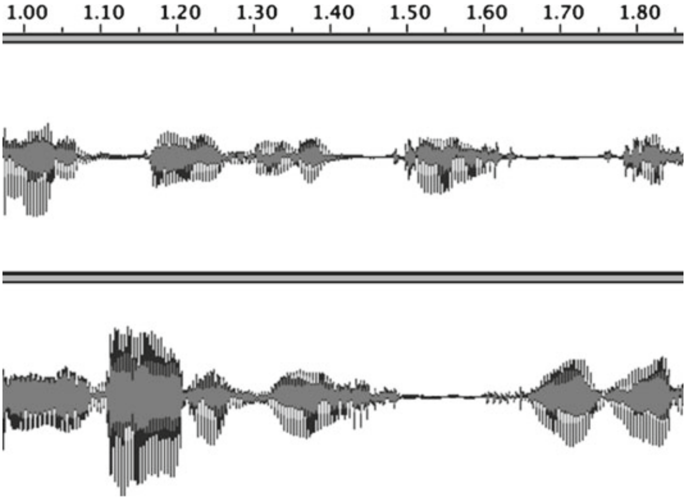


Fig. 16.5 Non-overlapping simultaneous speech

the simultaneous audio timeline, such results would be provided in a sporadic fashion. To this effect, a simple tracking system is proposed that dynamically clusters similar DOAs into candidate sound sources.

The tracker maintains in memory the last DOAs provided by the initial DOA estimator in a specific time frame. When a new DOA is estimated, the tracker carries out the following:

1. If the new DOA is not 'close enough' to the average DOA of any current cluster (good results were obtained when using 5° for clusters with one DOA, 10° for clusters with more than one DOA, as thresholds for closeness), or there are no clusters in the environment: create a new cluster with the new DOA.
2. If it is close enough to a current cluster, add the new DOA to it, and re-calculate its new average DOA.

If a DOA is too old (an age of 10 s provided good results), it is 'forgotten' by removing it from its respective cluster and re-calculating its average DOA.

Every cluster is considered a *candidate sound source*, until it has a pre-specified number of DOAs attributed to it (2 DOAs provided good and fast results), when it becomes a 'sound source' and its average DOA becomes its main estimated DOA.

16.4 Trials and Results

The test scenario was as follows: three microphones, 20 cm. apart from each other, were installed in the upper base of the Golem-II+ robot. In turn, it was placed in a large room with a high sonic complexity (considerably reflective materials, with a low ceiling, hard floor, cement columns in the middle, and moderate reverberation). Two electronic speakers emitting, simultaneously, random recordings from the DIMEX corpus [23] for 20 s, were placed at 1.5 m from the robot, one at 0°, another at $-45°$.

The Audio Acquisition module was sampling at 48 kHz, and providing sample windows of 4,800 samples (100 ms). The buffers in the VAD were 10 energy values long, and considering a 2 s delay for adjustment to the environment noise. The DOA estimator had a 40° incoherence threshold (any sample window set with a higher incoherence was rejected). The multi-DOA tracker considered a new DOA as part of a cluster with more than one DOA if it was 10° or closer to its average DOA; if the cluster only had one DOA, 5° or closer was considered as part of the cluster. The results of the test are shown in Fig. 16.6.

As it can be seen, the tracking system performed well with 2 sound sources (in the Figure referred to as 'Users').

The system was then tested with an additional simultaneous source: a human emitting continuously the phrase "golem i am over here (pause)" placed at 35°. The results of this scenario are shown in Fig. 16.7.

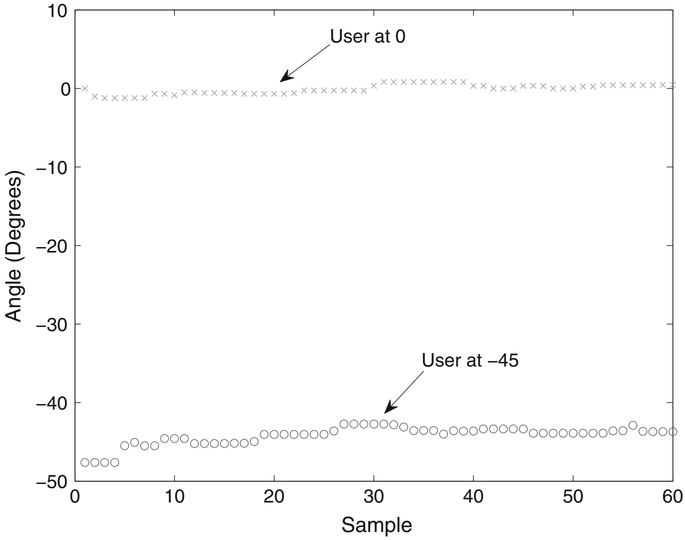


Fig. 16.6 Tracking 2 simultaneous sources (2 electronic speakers)

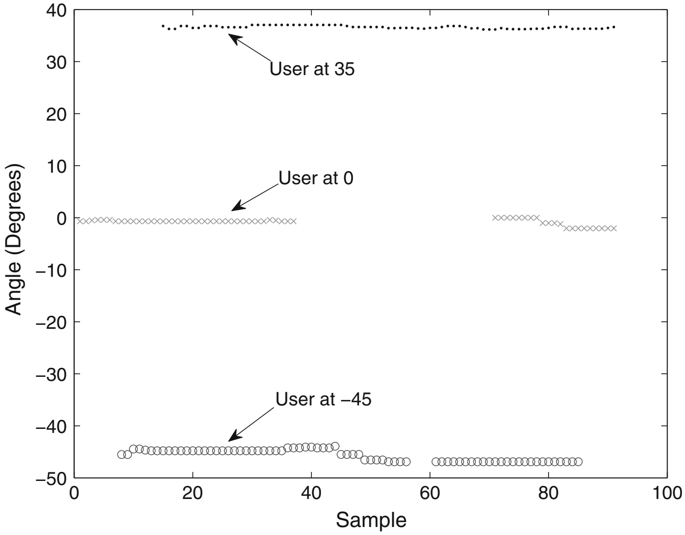


Fig. 16.7 Tracking 3 simultaneous sources (2 electronic speakers, 1 human)

As it can be seen, the system tracked the human and one of the electronic speakers (placed at −45°) well. The other of the two electronic speakers (placed at 0°) was 'missed' for a moderate amount of time, however, in any other moment, the tracking system was able to track it considerably well.

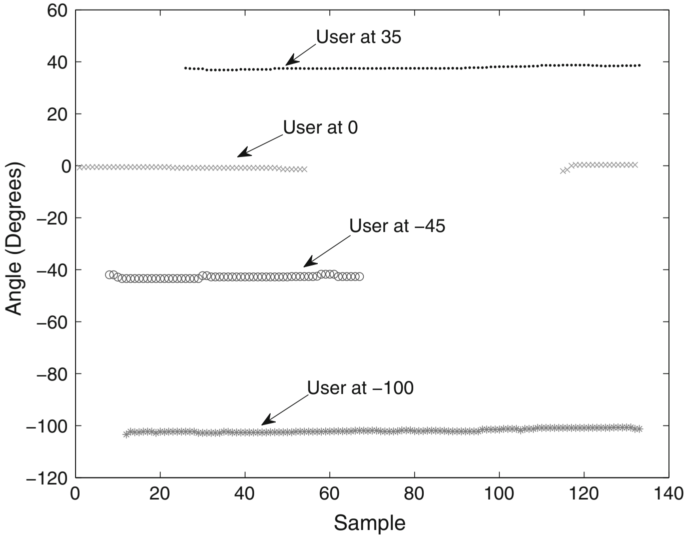


Fig. 16.8 Tracking 4 simultaneous sources (2 electronic speakers, 2 humans)

To assess if the ‘missed’ tracking issue was with the electronic speaker itself, and, in addition, to observe if the tracker is able to better identify humans than electronic speakers, an additional simultaneous source was added to the environment: another human emitting continuously the phrase “one two three (pause)” placed at −100°. The results of this final scenario are shown in Fig. 16.8.

And, as it can be seen, the electronic speaker placed at 0° was again ‘missed’ in a similar fashion than in the 3-user scenario, which suggests a failure with the specific characteristics of the electronic speaker (positioning, volume, frequency enveloping of speech, etc.). However, both humans were tracked very well, and the electronic speaker placed at −45° was tracked relatively well. These results imply that the proposed system is well suited for tracking simultaneous human speech.

In addition, and more significantly, for a moderate amount of time, the 4 simultaneous sources were being tracked well. Considering that the system only employs 3 microphones, it showed that it was able to monitor more sources than the number of microphones present, a feat that the popular approach known as MUSIC is unable to accomplish [16]. In fact, the number of sources that can be simultaneously tracked by the proposed system may not have a theoretical boundary, as speech overlap is the primary limiter, and, as previously stated, people do not tend to interrupt each other [21]. However, further testing is required to explore this topic.

The authors would like to remind the reader that the setting of the test scenario were considerably harsh: the sonic complexity of the room was high, there was moderate reverberation, the human user placement can be expected to be inconsistent, and no reverb adequation was carried out. When considering all of this, the proposed system has shown it is an adequate solution to the multi-user DOA estimation problem in a robotic mobile platform.

16.5 Conclusion and Future Work

Human-Robot Interaction benefits from a rich perception of the world. Having the robot orient itself towards the user during a conversation enhances HRI from the point of view of both the user and the robot: the 'naturality' of the conversation is improved, and the acquisition of more information from the user (face recognition, voice context, etc.) is simplified. To do this, however, the direction of the user is required. Because a conversation is carried out via voice, it is appropriate that the direction of the user be estimated by sound analysis.

In addition, multi-user scenarios are common in the day-to-day dynamics of a service robot. However, the multi-DOA estimation problem is further complicated when applied in mobile robotics, as it presents a unique challenge: the mobility of the robot should not be compromised, thus the hardware should be small and lightweight (limited amount of microphones), but it should be robust and flexible enough to be able to carry out DOA estimation in acoustically complex settings.

In this contribution, a 3-microphone system was proposed, built upon earlier work published by the authors. It provides a reliable Multiple Direction-of-Arrival estimation, and it was shown that it was able to track more users than the amount of microphones used. Moreover, it did so while being light enough to be carried by a service robot. It also provided a robust estimation in the presence of moderate reverberation and high sonic complexity.

However, during the evaluation, were human speech and electronic-speakers were emitting simultaneously, it was observed that the human speech overcame the electronic speakers. Although this might be attributed to specific characteristics of the hardware, it was observed that human speech was consistently tracked well, which is something desirable as it will be employed with real-life human speech.

This system is planned to be a preamble for a consequent module that will perform online source separation based on the DOA of the source, which will then provide the Automatic Speech Recognizer with speech data. This will result in a multiple-simultaenous-speech recognition, with a small hardware setup and redundant estimation.

Acknowledgments The authors would like to thank the support of CONACYT through the project 81965, and PAPIIT-UNAM with the project IN115710-3.

References

1. Lockwood ME, Jones DL, Bilger RC, Lansing CR, O'Brien WD Jr, Wheeler BC, Feng AS (2004) Performance of time- and frequency-domain binaural beamformers based on recorded signals from real rooms. J Acoust Soc Am 115(1):379–391
2. Hjelmås E, Low BK (2001) Face detection: a survey. Comput Vision Underst 83(3):236–274
3. Stiefelhagen R, Ekenel HK, Fugen C, Gieselmann P, Holzapfel H, Kraft F, Nickel K, Voit M, Waibel A (2007) Enabling multimodal human-robot interaction for the karlsruhe humanoid robot. IEEE Trans Robot 23(5):840–851

4. Faller C, Merimaa J (2004) Source localization in complex listening situations: selection of binaural cues based on interaural coherence. J Acoust Soc Am 116(5):3075–3089
5. Smith M, Kim K, Thompson D (2007) Noise source identification using microphone arrays. Proc Inst Acoust 29(5):8
6. Valin J, Rouat J, Michaud F (2004) Enhanced robot audition based on microphone array source separation with post-filter. In: Proceedings of IEEE/RSJ international conference intelligent robots and systems, pp 2123–2128
7. Murray JC, Erwin HR, Wermter S (2009) Robotic sound-source localisation architecture using cross-correlation and recurrent neural networks, What it means to communicate. Neural Networks 22(2): 173–189
8. Pineda L, Meza I, Aviles H, Gershenson C, Rascon C, Alvarado-Gonzalez M, Salinas L (2011) Ioca: interaction-oriented cognitive architecture. Res Comput Sci 54:273–284
9. Liu R, Wang Y (2010) Azimuthal source localization using interaural coherence in a robotic dog: modeling and application. Robotica First View, pp 1–8
10. Horchler AD, Reeve RE, Webb B, Quinn RD (2003) Robot phonotaxis in the wild: a biologically inspired approach to outdoor sound localization. In: Sound localization 11 th international conference on advanced robotics, (ICAR ’03), pp 1749–1756
11. Murray JC, Erwin H, Wermter S (2004) Robotics sound-source localization and tracking using interaural time difference and cross-correlation. AI Workshop on NeuroBotics
12. Nakadai K, Okuno HG, Kitano H (2002) Real-time sound source localization and separation for robot audition. In: Proceedings IEEE international conference on spoken language processing, 2002, pp 193–196
13. Wang D, Brown GJ (eds) (2006) Computational auditory scene analysis: principles, algorithms, and applications. IEEE Press/Wiley-Interscience. URL http://www.casabook.org/
14. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. Robot Auton Syst 42(3–4):143–166
15. Saxena A, Ng AY (2009) Learning sound location from a single microphone. In: ICRA’09: proceedings of the 2009 IEEE international conference on robotics and automation, pp 4310–4315. IEEE Press, Piscataway, NJ, USA
16. Schmidt R (1986) Multiple emitter location and signal parameter estimation. Antennas Propag IEEE Trans 34(3):276–280
17. Mohan S, Lockwood ME, Kramer ML, Jones DL (2008) Localization of multiple acoustic sources with small arrays using a coherence test. J Acoust Soc Am 123(4):2136–2147
18. Rascon C, Pineda L (2012) Lightweight multi-doa estimation on a mobile robotic platform. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct, San Francisco, USA, pp 665–670
19. Rascon C, Aviles H, Pineda LA (2010) Robotic orientation towards speaker for human-robot interaction. Adv Artif Intell IBERAMIA 6433:10–19
20. Davis P (2001) Jack, connecting a world of audio. http://jackaudio.org/(2001)
21. Stivers T, Enfield NJ, Brown P, Englert C, Hayashi M, Heinemann T, Hoymann G, Rossano F, de Ruiter JP, Yoon KE, Levinson SC (2009) Universals and cultural variation in turn-taking in conversation. Proc Natl Acad Sci 106(26):10587–10592
22. Shriberg E, Stolcke A, Baron D (2001) Observations on overlap:Findings and implications for automatic processing of multi-party conversation. Proc Eurospeech 2:1359–1362
23. Pineda L, Castellanos H, Cuetara J, Galescu L, Juarez J, Listerri J, Perez P, Villaseñor L (2010) The corpus dimex100: transcription and evaluation. Lang Resour Eval 44(4):347–370

Chapter 17
Analysis of Metallic Plume Image Characteristics During High Power Disk Laser Welding

Xiangdong Gao, Runlin Wang, Yingying Liu and Yongchen Yang

Abstract Metallic plume is an important phenomenon during high power disk laser deep-penetration welding, which can reflect the welding quality. To study this laser-induced plume characteristics and its relation to welding quality, an extraviolet and visible sensitive high speed color camera was used to capture the metallic plumes in a high-power disk laser bead on plate deep-penetration welding of Type 304 austenitic stainless steel plates at a continuous laser power of 10 kW. These captured digital images were firstly processed in RGB color spaces, and then were transferred to the Hue-Saturation-Intensity (HSI) color spaces from the RGB color spaces. The area of metallic plume was segmented and defined as the plume eigenvalue. The fluctuation of weld bead width was used to evaluate the welding stability. To monitor the plume behavior, a short-time Fourier transform was applied to obtain the time–frequency characteristics of plume images. Also, the hierarchical clustering was analyzed for the time–frequency characteristics of plume images. The results of hierarchical clustering showed there existed relationship between the metallic plume area and welding quality, and the fitting curve of clustering could reflect the fluctuation trend of the weld bead width effectively.

Keywords Disk laser welding · Hierarchical clustering · Image characteristics · Metallic plume image · Short-time Fourier transform · Time–frequency analysis

X. Gao (✉) · R. Wang · Y. Liu · Y. Yang
School of Electromechanical Engineering, Guangdong University of Technology,
510006 Guangzhou, China
e-mail: gaoxd666@126.com

R. Wang
e-mail: xdgao@hotmail.com

Y. Liu
e-mail: liu15045105965@126.com

Y. Yang
e-mail: gaoxd@gdut.edu.cn

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_17,

17.1 Introduction

High power disk laser welding is a competitive welding method and is well known for its high welding speed, good welding quality and deep penetration. In recent years, the disk laser welding has been widely used in automotive production and electronic industry. During a high power disk laser welding, a metallic plume mixture is generated quickly from the surface of the welded material. This plume mainly consists of the metal vapor and is one of the most important phenomenon which can be used to monitor the laser welding quality. Research works have shown that the metallic plume has negative effects on the energy transference efficiency of the laser beam and the welding quality [1, 2]. There exists an internal relationship between the plume characteristics and the welding status. To monitor and control the welding quality in real-time, it is necessary to investigate the metallic plume characteristics and its influences upon the weld quality.

In recent years, some researches such as spectroscopy, photoelectric signal processing, acoustic signal processing, vision methods, and so on were performed to study the dynamic behaviors of the metallic plume [3–6]. The holographic interferometry was applied to study the laser-induced plume [7]. The Fourier transform was used to analyze the acoustic signal of metal vapor and the time–frequency was applied to study the plasma characteristics [8, 9]. These study results showed that the density and the size of the plasma were related to the laser power and beam focus position. There were relations between the metallic plume and the weld quality. However, it is still difficult to find the exact relationship between the characteristics of metallic plume and the weld quality.

In order to obtain more detailed dynamic information about metallic plume, the high-speed photography was used to record the color images of metallic plume in a high power disk laser welding process. High-speed photography is an effective method and is widely used in the welding process, it can accurately capture and monitor the instant information of metallic plume. The important characteristic features of plumes could be extracted from these color images [10, 11]. Usually, the more energy a weldment absorbs, the bigger the metallic plume is. Here, the metallic plume area was used as the characteristic parameters, and the short-time Fourier transform was applied to the area of metallic plume to obtain the time–frequency characteristics of plume. Also, the hierarchical clustering was used to analyze the plume characteristics and finally the clustering curves of metallic plume area were plotted. Welding experimental results showed that in a definite parameter combination, the 6th fitting curves of the metallic plume area frequency characteristic clustering could effectively reflect the fluctuation trend of the weld bead width.

17.2 Experimental Apparatus and Plume Characteristic Extraction

17.2.1 Experimental Apparatus

The schematic of a disk laser welding experimental apparatus is shown in Fig. 17.1. The experimental system consisted of a TruDisk-10003 disk laser welding equipment (laser power 10 kW), a Motoman 6-axis robot and a welding experimental platform equipped with shielding gas (argon), servo motors and fixing devices. An extraviolet and visible sensitive high speed color camera was used to capture the metallic plume dynamic color images during a 10 kW high-power bead-on-plate disk laser welding. The welding conditions are listed in Table 17.1

17.2.2 Extraction of Plume Characteristics in RGB Space

The high-speed camera collected 2,400 frames RGB image of the metallic plume within 1.2 s, and each frame image corresponded to a welding status. The top view of a welded specimen is shown in Fig. 17.2. It can be seen that the middle part of

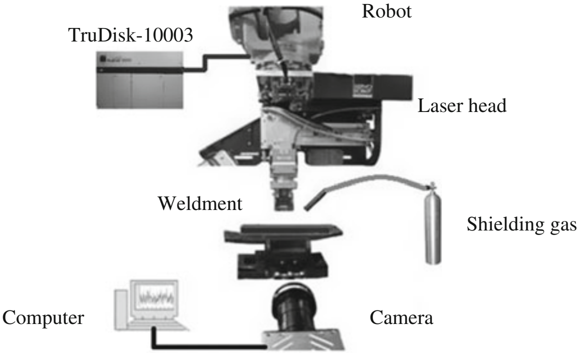


Fig. 17.1 Experimental apparatus of high power disk laser welding

Table 17.1 Welding experimental conditions

Welding apparatus	TruDisk-10003
Laser power	10 kW
Spot diameter	480 μm
Laser wavelength	1,030 nm
Welding speed	4.5 m/min
Camera speed	2,000 frame/s
Image resolution	512 × 512 pixel
Size of weldment	150 × 100 × 10 mm
Weldment	Type 304 austenitic stainless steel

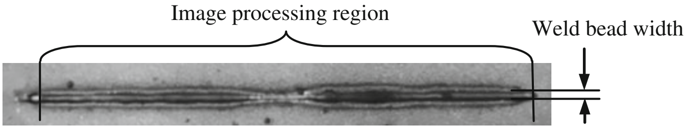
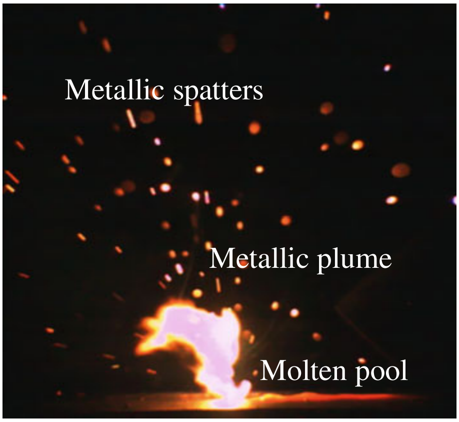


Fig. 17.2 *Top* view of a welded specimen of high-power disk laser welding

Fig. 17.3 Original metallic plume color image



the weld seam is narrow and has poor quality. This region corresponded to 1,066–1,333 frame images. The captured plume images from 481 to 2,400 frames were processed to study their characteristics. An original metallic plume color image is shown in Fig. 17.3. It can be seen that the captured image includes the information of metallic spatters, metallic plume and molten pool.

When the disk laser beam focused on a weldment, the laser energy was transfered to the surface of weldment, the weldment melted immediately and the metallic plume emerged. The area of metallic plume could reflect the absorptivity of laser energy which reflected the welding quality. Thus, the metallic plume area could be used as a characteristic parameter. In order to extract the plume characteristics accurately and reduce calculation, metallic plume were tailored from an original color image. The tailored RGB image is shown in Fig. 17.4.

The tailored plume image was turned to gray scale image. Due to random interference signals in the welding, there were a lot of noises in the image when it was recorded, the filtering was used to eliminate these noises. Commonly used filtering methods are frequency filtering and spatial filtering. This experiment used the spatial filtering to deal with the noises. Spatial filtering can be divided into linear filtering and nonlinear filtering. Linear filtering is also called as mean filtering, and is a low pass filter. Because the profile of image edge contains a lot of high frequency information, so the boundary of image becomes fuzzy by using the mean filtering eliminating the noises. Boundary is one of the most basic image features and often carries much image information that is of great importance in analyzing, describing and understanding an image.

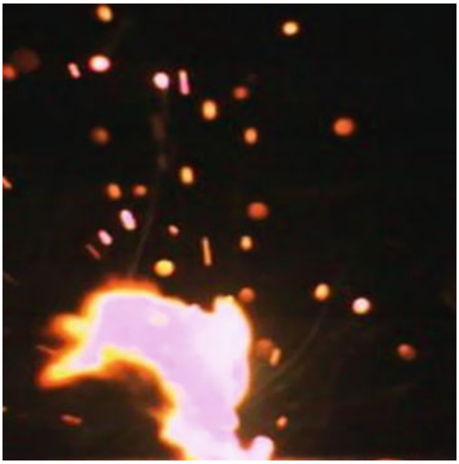


Fig. 17.4 Tailored *RGB* image of metallic plume

Wiener filtering and Median filtering are commonly used for nonlinear filtering and they can not only filter all kinds of noises, but also protect the boundary information such as edge and sharp corner. Wiener filtering characterized by the feature of good recovery effect, low computation and good performance of noise reduction, are widely used in image recovery. Wiener filtering was used in this experiment. The experiment applied a group of filtering windows to conduct the Wiener filtering processing, and the results are shown in Fig. 17.5. It is observed from Fig. 17.5 that the filtering effect of 35 × 35 filtering window is best. It also can be seen from Fig. 17.5c that the noises such as spatter, halo and so on were filtered by Wiener filtering, the edge information was protected and the images were clear.

Image segmentation is of importance in image processing. The accuracy of image segmentation has a direct influence on the subsequent image refining and recognition results. Image segmentation is based on the edge, shape, gray value and position to divide an image into different kinds of regions, and separates the target image from background. The commonly used method of image segmentation is threshold segmentation such as image binarization. When the gray scale of target is greatly different from background and the layers of image are clear, threshold segmentation can better detect the target. Threshold segmentation can not only compress the data quantity, but also simplify the analysis and processing steps. Therefore, in many cases, it is an indispensable image preprocessing process, feature extraction and pattern recognition. Threshold segmentation method principle can be described as follows. Lets an original grayscale image is $f(x,y)$, then a gray value T of original grayscale image $f(x,y)$ can be found with some certain criteria to divide image into two parts. The divided binary image is given by

$$g(x,y) = \begin{cases} a & f(x,y) \geq T \\ b & f(x,y) < T \end{cases} \tag{17.1}$$

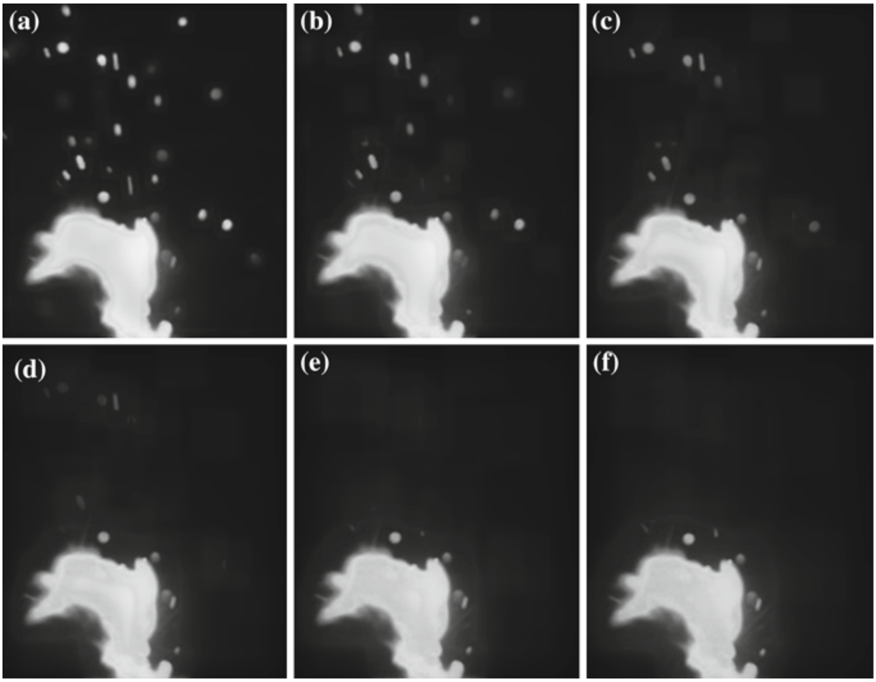


Fig. 17.5 Effect diagram of Wiener filter. **a** 15 × 15 filtering window. **b** 25 × 25 filtering window. **c** 35 × 35 filtering window. **d** 45 × 45 filtering window. **e** 55 × 55 filtering window. **f** 65 × 65 filtering window

The image belonging to different target region is defined by threshold value, so the selection of optimal threshold *T* is the key for deciding the effect of threshold segmentation. If the threshold value is too high, overmuch target points are classified as the background falsely. If the threshold value is too low, the opposite situation appears. Commonly used threshold selection method is P—tile method, Otsu method, average gray method, the waterline threshold method, the maximum entropy method and fuzzy set method and so on. Owing to the complex algorithm, the latest methods like the maximum entropy method and fuzzy set method are not applicable to this experiment. P—tile method, Otsu method, average gray method and the waterline threshold method are applied and the results are shown in Fig. 17.6. It can be observed from Fig. 17.6 that the size and morphology of the metallic plume is the same with the Otsu image and it could meet the computing requirement. After image segmentation, there were a lot of spatters in some metallic plume images, and this would influence the extraction of plume characteristics, so the image areas with less than 200 pixels were deleted and the final metallic plume images could be obtained.

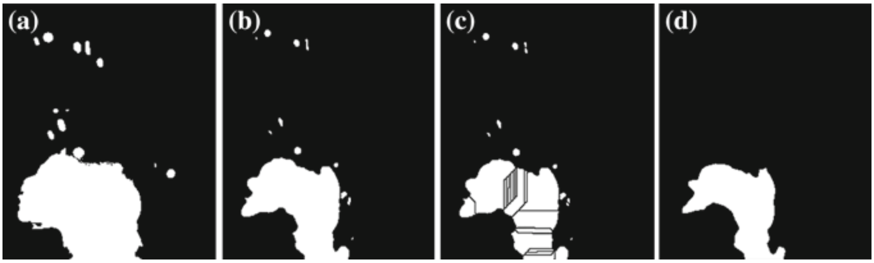


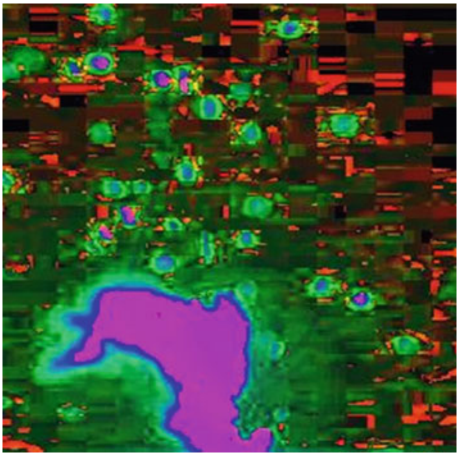
Fig. 17.6 Image segmentation of laser-induced metallic plumes. **a** Average gray method. **b** P—tile method. **c** Waterline threshold method. **d** Otsu method

17.2.3 Extraction of Plume Characteristics in HSI Space

Color space has many description including RGB color spaces, HIS color space plays an important role in image analysis. HSI color space uses the hue, saturation and intensity to describe colors which have a better performance in scenery cognitive than RGB color spaces. Here we used HSI method to process the plume images. An original RGB image of metallic plume shown in Fig. 17.4 was tailored. Then this tailored RGB image of metallic plume was converted to the HSI image and is shown in Fig. 17.7. Respectively, *H* component, *S* component and *I* component could be obtained from the mathematical formulas. *H* component is given by

$$H = \begin{cases} \theta & if \quad B \le G \\ 360 - \theta & if \quad B > G \end{cases} \tag{17.2}$$

Fig. 17.7 HSI image of metallic plume



where

$$\theta = \arccos\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{1/2}}\right\} \tag{17.3}$$

S component is given by

$$S = 1 - \frac{3}{(R+G+B)}[\min(R,G,B)] \tag{17.4}$$

I component is given by

$$I = \frac{1}{3}(R,G,B) \tag{17.5}$$

where *R*, *G*, *B* are three components of RGB color spaces.

H component image, *S* component image and *I* component image are shown in Fig. 17.8. The central part of metallic plume absorbs more laser energy and has a key influence on laser beam and welding. *H* component reflects the central part of metallic plume which can be used to extract the plume characteristics. In the *H* component image, the body of metallic plume is obviously different from black background. The Otsu threshold segmentation was used to segment the metallic plume. Setting global threshold 200 could remove the white spots and the metallic plume was segmented. It can be observed that the segmented metallic plume contains several small black holes. Through filling these holes, the final metallic plume is achieved. The image processing procedure of metallic plume is shown in Fig. 17.9. The plume area of whole metallic plume images were calculated, as shown in Fig. 17.10.

The metallic plume area was used to study the plume characteristics. In this experiment, the image processing methods of both RGB spaces and HSI spaces were available, and from the curves it is difficult to find there existed the obvious fluctuations of metallic plume area. Therefore, we considered applying the

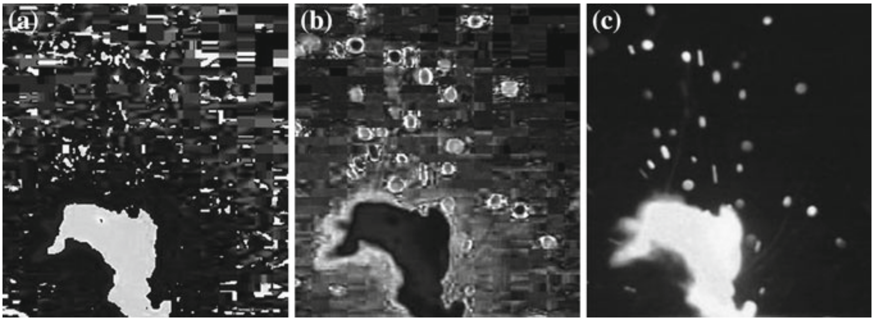


Fig. 17.8 **a** *H* component image. **b** *S* component image and **c** *I* component image of metallic plume

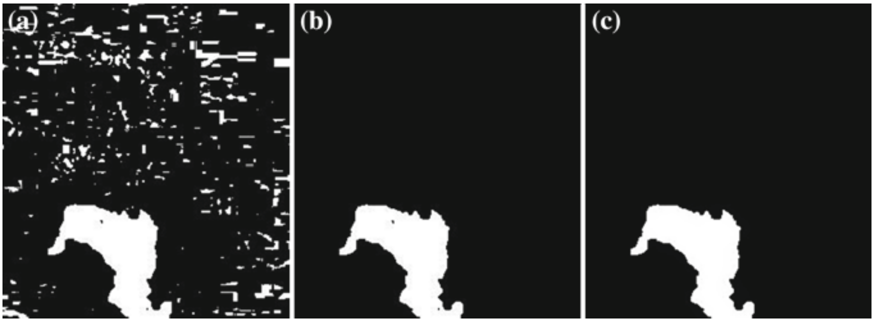
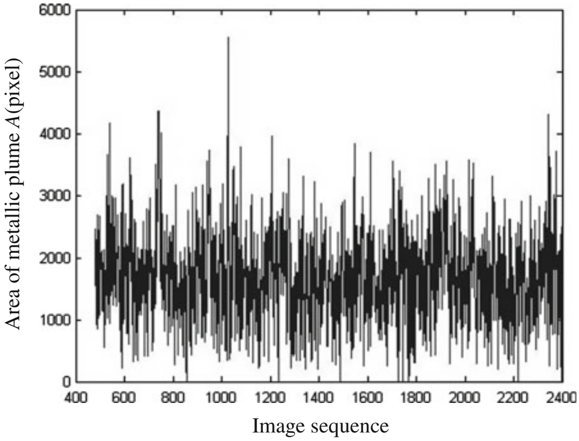


Fig. 17.9 Schematic diagram of metallic plume image processing. **a** Binary image. **b** Segmented image. **c** Final image

Fig. 17.10 *Curves* of metallic plume area with Image sequences



methods of short-time Fourier transform and hierarchical clustering to investigate the plume characteristics in HSI color spaces.

17.3 Short-Time Fourier Transform

17.3.1 Concept of Short-Time Fourier Transform

Short-time Fourier transform (STFT) can not only reflect the time-domain feature of signals, but also present the spectrum of signals clearly. Its basic idea is that the signal to be transformed is multiplied by a limited window function before the Fourier transform is applied, and this window function is nonzero for only a short period of time. This window slides along the time axis, resulting in a two-dimensional representation of the signal. This can be mathematically written as [12]

$$STFT_Z(t,f) = \int_{-\infty}^{\infty} z(t)\eta^*(t - t')e^{-j2\pi ft}dt \tag{17.6}$$

where $z(t)$ is the signal to be transformed and $\eta^*(t - t')$ is the window function around t'. Through $z(t)\eta^*(t - t')$, the signal around t' is obtained and the short-time Fourier transform is just the Fourier transform of $z(t)\eta^*(t - t')$.

17.3.2 Window Function

The frequently-used window functions are Rectangular window, Gauss window, Hanning window, Hamming window, Blackman window, Triangle window, Cosine slope window, Index window and Bartlett-Hanning window. In welding experiment, Gauss window, Hanning window, Hamming window and Bartlett-Hanning window were applied to the short-time Fourier transform.

Suppose $x(n)$ is the signal sequence and $w(n)$ is a window function whose length is N. The expression of the Gauss window is

$$w(n) = e^{-\frac{1}{2}\left(\frac{n-(N-1)/2}{\sigma(N-1)/2}\right)^2} \tag{17.7}$$

where $\sigma \leq 0.5$.

The expression of the Hanning window is

$$w(n) = 0.5 - \left(1 - \cos\left(\frac{2\pi n}{N - 1}\right)\right) \tag{17.8}$$

The expression of the Hamming window is

$$w(n) = 0.53 - 0.46\cos\left(\frac{2\pi n}{N - 1}\right) \tag{17.9}$$

The expression of the Bartlett-Hanning window is

$$w(n) = 0.42 - 0.5\cos\left(\frac{2\pi n}{N - 1}\right) + 0.08\cos\left(\frac{4\pi n}{N - 1}\right) \tag{17.10}$$

In order to improve the temporal resolution of short-time Fourier transform, the length of window function should be as short as possible. At the same time, in order to get a higher frequency resolution, the length of the window should be as long as possible. In practical application, the length of the window function should be adapted to the length of signal local smooth length [13]. In laser welding experiment, the numerical values of length were set to be 64, 128, 256, respectively.

17.3.3 Analysis of Short-Time Fourier Transform

In short-time Fourier transform, the different window types, different window length and different step length were chosen. The detailed combination parameters are listed in Table 17.2.

Using the short-time Fourier transform, $4 \times 3 \times 3 = 36$ groups of data were obtained, in which the numbers of window types, length, step length were 4, 3, 3, respectively. Taking a group of data for example, the parameters were Gauss, 64, 10. The window length was 64 and it slid along the time axis 186 times during the short-time Fourier transform, so this group of data was a matrix whose size was 64×186. Figure 17.11 is a 3-D map of time–frequency information and Fig. 17.12 is the contour map of time–frequency.

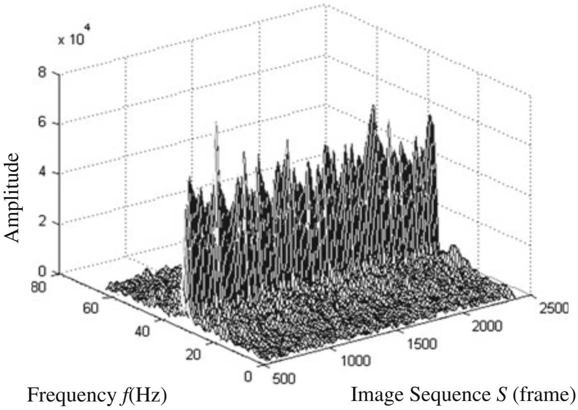
As mentioned above, the image sequence 1,066–1,333 frames corresponded to the middle part of the weld bead. This region of weld bead was narrow and had poor quality. Observing Figs. 17.11 and 17.12, there were not obvious characteristics of 1,066–1,333 frames. For further study, the 50th, 70th, 110th frequency curves were extracted to analyze their characteristics. These three groups of data corresponded to three vertical lines, shown in Fig. 17.12. Figure 17.13 shows these three frequency curves.

To distinguish these three frequency curves more effectively, their numerical values of average, maximum, minimum, range, interquartile range (IQR), standard deviation and sum were calculated. The range was what the biggest number minus the smallest number. The IQR is the distance between the 75th percentile and the 25th percentile. The expression of standard deviation is

Table 17.2 Combination parameters of short-time Fourier transform

Window types	Gauss, Hanning, Hamming, Bartlett-Hanning
Length	64, 128, 256
Step length	1, 5, 10

Fig. 17.11 3-D Map of time–frequency of plume area. Frequency *f*(Hz)



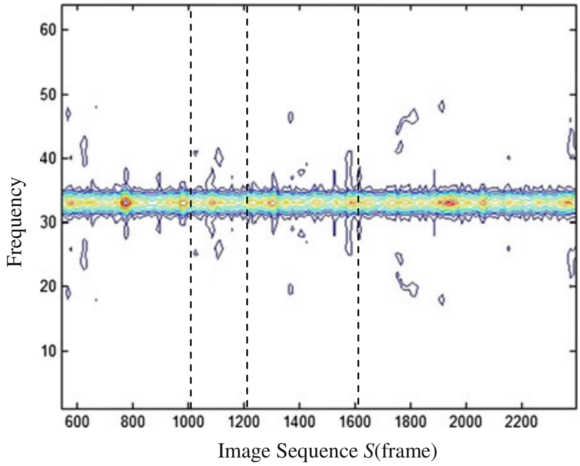


Fig. 17.12 Contour map of time–frequency of plume area

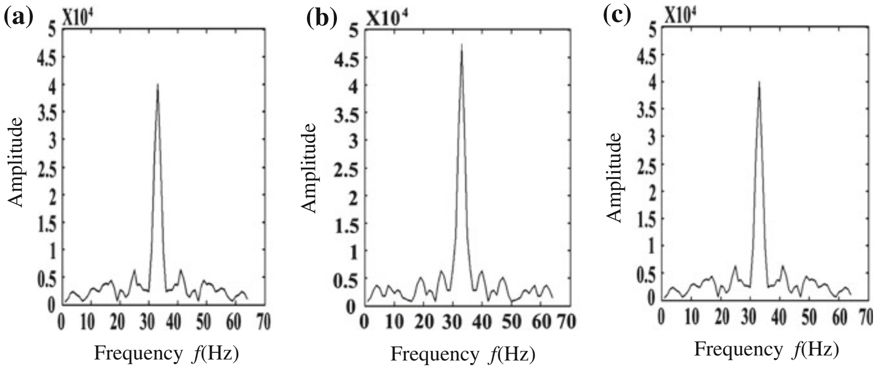


Fig. 17.13 Curves of the 50th, 70th, 110th frequency of plume area signals. **a** The 50th group of data. **b** The 70th group of data. **c** The 110th group of data

$$s = \left(\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^{1/2} \qquad i = 1, 2, 3 \ldots n \tag{17.11}$$

where $\bar{x}$ is the average value. All these eigenvalues are listed in Table 17.3.

It was found from Fig. 17.13 that three frequency curves had similar shapes. These three curves could be distinguished from Table 17.3 effectively by seven eigenvalues. Thus, these seven eigenvalues could represent different spectrum curves at any time. Also, we used the statistical method to calculate all frequency curve eigenvalues and analyze them by the Hierarchical clustering.

Table 17.3 Eigenvalues of different curves among Fig. 17.13 (unit:a.u)

	Fig. 17.13a	Fig. 17.13b	Fig. 17.13c
Max value	39,922	47,275	42,242
Min value	573	815	107
Average	4,199	4,840	3,787
IQR	1,893	2,318	2,825
Range	39,348	46,459	42,134
Standard deviation	6,435	7,702	6,743
Sum	268,750	309,800	242,410

17.4 Hierarchical Clustering

With the development of multivariate statistic analysis, the clustering analysis method has been mature gradually and widely used in Biology, Economics, Sociology, Demography and so on. The hierarchical clustering is the most important method in clustering analysis. Its basic principle is that the two closest observations are joined to create a node by calculating the distance or similar coefficient between two observations. Subsequent nodes are created by pairwise joining of observations or nodes based on the distance between them, until all the nodes merge into a desired number of clusters. At the end, a tree structure can be created by retracing which items and nodes are merged [14].

In order to decide which clusters should be combined or where a cluster should be split, a measurement of dissimilarity between sets of observations is required. In most methods of hierarchical clustering, this can be achieved by using an appropriate metric (a measure of distance between pairs of observations) and a linkage criterion which specifies the dissimilarity of sets as a function of the pairwise distances of observations.

Some commonly used distance metrics for hierarchical clustering are the Euclid distance, Minkowski distance, City Block distance, Chebyshev distance, Mahal distance, Lance distance and Cosine similarity. The linkage criteria determines the distance between sets of observations as a function of the pairwise distances between observations. There are a variety of linkage criteria between clusters. Among them, three most popular ones are maximum or complete linkage, minimum or single linkage, mean or average linkage [15]. In our welding experiments, we defined the Euclid distance and the City Block distance as the distance metrics and took minimum linkage for hierarchical clustering. The expression of Euclid distance is

$$d_{ij}^{(2)} = \left(\sum_{t=1}^{p} \left| x_{it} - x_{jt} \right|^2 \right)^{1/2} \tag{17.12}$$

The expression of City Block distance is

$$d_{ij} = \left(\sum_{t=1}^{p} \left| x_{it} - x_{jt} \right| \right) \tag{17.13}$$

Mathematically, the minimum linkage is written as

$$D_{pq} = \min(d_{ij}) \quad x_i \in G_p, \quad x_j \in G_q \tag{17.14}$$

After the short-time Fourier transform, there were 36 groups of time frequency test data. The statistics method was used to extract the seven defined eigenvalues, then these eigenvalues were studied by clustering. The procedure of hierarchical clustering is as follows. First, the Euclid distance and the City Block distance were defined as the distance metrics, and the distance between observations were calculated. Second, the minimum linkage was used to create a tree structure. Finally, the discontinuous coefficients was set and the clustering tree was output. In this experiment, the discontinuous coefficients were 0.5, 0.7 and 0.9, respectively.

Totally 36 × 2 × 3 = 216 groups of clustering data were obtained after calculation. The characteristics of clustering data was analyzed by drawing the clustering curves and curve-fittings. Using Bartlett-Hanning window whose length was 256 and step value was 5 for the short-time Fourier transform, and the Euclid distance, minimum linkage and discontinuous coefficient 0.9 for hierarchical cluster, it was found that the fluctuation trend of 6th fitting of clustering curve was similar to the 6th fitting curve of weld seam bead width. That means the 6th fitting of the clustering curve could reflect the weld bead width changing trend effectively. Figure 17.14 is the clustering curve based on the combination parameters mentioned above, the dotted line is the 6th fitting curve of clustering curve.

Figure 17.15 shows a 6th fitting curve of the clustering curve contrasting to the 6th fitting curve of weld seam bead. In Fig. 17.15, the dotted line is the weld bead width, the dot and dash line is the 6th fitting curve of weld bead width and the solid line is the 6th fitting curve of the clustering curve. It can be seen that the solid line and the dot and dash line have the consistent fluctuations. The 6th fitting curve of the clustering curve could reflect the weld bead width change trend effectively.

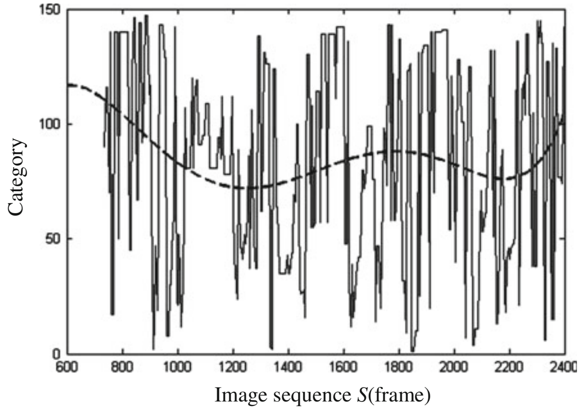


Fig. 17.14 Fitting *curve* of time–frequency clustering of metallic plume area

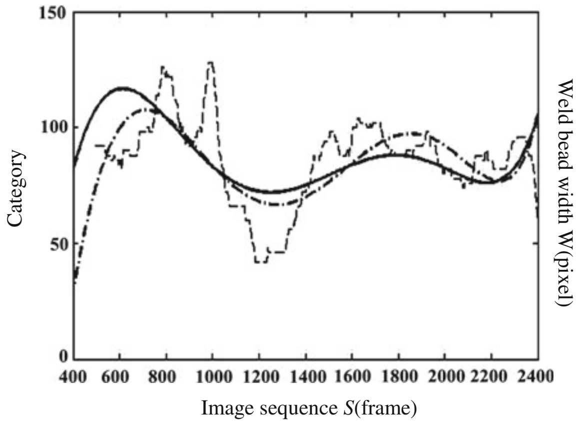


Fig. 17.15 Description of time–frequency clustering of area and weld bead width

Experimental results have shown that the weld bead width can be evaluated by using the time–frequency clustering of metallic plume area. It has provided a method to monitor and evaluate the welding quality in real time during disk laser welding by analyzing the time–frequency clustering of metallic plume area.

17.5 Conclusions

In a high power disk laser welding process, there exists a relation between the metallic plume area and the weld bead width. The metallic plume area could be calculated by using image processing techniques. It was found that the accurate plume area could be obtained by processing the plume images in the RGB color space and HSI color space.

The short-time Fourier transform could be applied to analyze the characteristics of plume area and extract the eigenvalues for the hierarchical clustering. Using the Bartlett-Hanning window whose length was 256 and step value was 5 for the short-time Fourier transform, and the Euclid distance, minimum linkage and discontinuous coefficient 0.9 for the hierarchical clustering, the 6th fitting curve of the clustering curve and the 6th fitting curve of weld bead width had the similar fluctuations. The 6th fitting of the clustering curve could reflect the weld bead width changing trend effectively. Experimental results showed that the time–frequency clustering of metallic plume area could be used to monitor and evaluate the welding quality during high power disk laser welding.

Acknowledgments This work is an expanded version of the paper published at WCECS 2012 in San Francisco, USA, October 24–26, 2012, and was supported in part by the National Natural Science Foundation of China under Grant 51175095, in part by the Guangdong Provincial Natural Science Foundation of China under Grants 10251009001000001 and 9151009001000020, and in part by the Specialized Research Fund for the Doctoral Program of Higher Education under Grant 20104420110001. Many thanks are given to Katayama Laboratory, Osaka University, Japan, for their assistance of laser welding experiments.

References

1. Khan MMA, Romoli L, Fiaschi M, Dini G, Sarri F (2011) Experimental design approach to the process parameter optimization for laser welding of martensitic stainless steels in a constrained overlap configuration. Opt Laser Technol 43(1):158–172
2. Katayama S, Kawahito Y, Mizutani M (2010) Elucidation of laser welding phenomena and factors affecting weld penetration and welding defects. Phys Procedia 5(Part B):9–17
3. Liu L, Chen M (2011) Interactions between laser and arc plasma during laser–arc hybrid welding of magnesium alloy. Opt Lasers Eng 49(9–10):1224–1231
4. Gao XD, Wang RL, Long GF, Katayama S (2012) Study of characteristics of plume based on hue-saturation-intensity during high-power disk laser welding. ACTA Physica Sinica, 61(14):148103-1-8
5. Khaleeq-ur-Rahman M, Siraj K, Rafique MS, Bhatti KA, Latif A, Jamil H, Basit M (2009) Laser induced plasma plume imaging and surface morphology of silicon. Nuclear Instruments Methods Phys Res B 267(7):1085–1088
6. Sibillano T, Anocona A, Berdi V, Lugara PM (2005) Correlation analysis in laser welding plasma. Opt Commun 251(1–3):139–148
7. Baik SH, Park SK, Kim CJ, Kim SY (2001) Holographic visualization of laser-induced plume in plused laser welding. Opt Laser Technol 33(1):67–70
8. Jiang P, Chen WZ, Guo J, Tian ZL (2001) The FFT Analyze of the acoustic signal on plasma in laser welding. Laser J 22(5):62–63
9. Molino A, Martina M, Vacca F, Masera G, Terreno A, Pasquettaz G, Angelo G (2009) FPGA implementation of time-frequency analysis algorithms for laser welding monitoring. Microprocessors Microsystems 33(3):179–190
10. Li G, Cai Y, WU Y (2009) Stability information in plasma image of high-power CO2 laser welding. Opt Lasers Eng 47(9):990–994
11. Gao XD, Wang RL, Yang YC (2012) Time-frequency characteristics clustering of metallic plume during high power disk laser welding, lecture notes in engineering and computer science. In: Proceedings of The world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct 2012, San Francisco, USA, pp 660–664
12. Wang X, Cheng P, Liang J (2011) Research of STFT time-frequency analysis algorithm and its application in train vibration analysis. Noise Vibr Control 31(1):65–68
13. Ge Z, Chen Z (2006) MATLAB time-frequency analysis technology and its application. Posts and Telecom Press, Beijing, pp 1–8
14. Du Z, Lin F (2004) A hierarchical clustering algorithm for MIMD architecture. Comput Biol Chem 28(5–6):417–419
15. Zhang Q, Zhang Y (2006) Hierarchical clustering of gene expression profiles with graphics hardware acceleration. Pattern Recogn Lett 27(6):676–681

Chapter 18
Accurate Spectral Estimation of Non-periodic Signals Based on Compressive Sensing

Isabel M. P. Duarte, José M. N. Vieira, Paulo J. S. G. Ferreira and Daniel Albuquerque

Abstract In this work we propose a method based on compressive sensing (CS) for estimating the spectrum of a signal written as a linear combination of a small number of sinusoids. In practice one deals with signals with finite-length and so the Fourier coefficients are not exactly sparse. Due to the leakage effect in the case where the frequency is not a multiple of the fundamental frequency of the DFT, the success of the traditional CS algorithms is limited. To overcome this problem our algorithm transform the DFT basis into a frame with a larger number of vectors, by inserting a small number of columns between some of the initial ones. The algorithm takes advantage of the compactness of the interpolation function that results from the ℓ_1 norm minimization of the Basis Pursuit (BP) and is based on the compressive sensing theory that allows us to acquire and represent sparse and compressible signals, using a much lower sampling rate than the Nyquist rate. Our method allow us to estimate the sinusoids amplitude, phase and frequency.

Keywords Basis Pursuit · Compressive sensing · Interpolating function · Redundant frames · Sparse representations · Spectral estimation

I. M. P. Duarte (✉)
School of Technology and Management of Viseu, Polytechnic Institute of Viseu and Signal Processing Lab., IEETA/DETI, University of Aveiro, Aveiro, Portugal
e-mail: isabelduarte@estv.ipv.pt

J. M. N. Vieira · P. J. S. G. Ferreira · D. Albuquerque
Signal Processing Laboratory, IEETA/DETI, University of Aveiro, Aveiro, Portugal
e-mail: jnvieira@ua.pt

P. J. S. G. Ferreira
e-mail: pjf@ua.pt

D. Albuquerque
e-mail: dfa@ua.pt

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_18,

18.1 Introduction

The compressive sensing (CS) theory allows us to recover, sparse or compressible signals, from a number of measurements M, much smaller than the length N of the signal. Instead of acquiring N samples, compute all the transform coefficients, discard the small $(N-K)$ and then encode the largest K, we can acquire a number of random mixtures M proportional to the sparsity K. In CS we acquire and compress the signal in one step.

The samples are obtained projecting x on a set of M vectors $\{\Phi_i\} \in \mathbb{R}^{\mathbb{N}}$, that are independent of the signal, with which we can build the measurement matrix $\Phi \in \mathbb{R}^{\mathbb{M}\times\mathbb{N}}$, with $M < N$. In matrix notation, the measurement vector $y = \Phi x$.

To reconstruct the K-sparse signal x, we search for the sparsest coefficient vector x, solving the problem:

$$(P0) : \min \|x\|_0 : y = \Phi x, \tag{18.1}$$

where $\|x\|_0$ is the number of nonzero entries. This problem is combinatorial, and so, to avoid this difficult we must use other approach. Since the matrix Φ is rank deficient, and so it loses information, one can think the problem is impossible, but it can be shown that if the matrix obeys the Restricted Isometry Property (RIP), we can recover x exactly by solving the convex problem, which is the convex relaxation of (P0) [1, 2]:

$$(P1) : \min \|x\|_1 : y = \Phi x, \tag{18.2}$$

where $\|x\|_1 = \Sigma|x_i|$.

Essentially, the RIP requires that every set of less than K columns, approximately behaves like an orthonormal system. More precisely, let Φ_T, $T \subset \{1, \cdots, N\}$, be the $M \times |T|$ submatrix consisting of the columns indexed by T. The K-restricted isometry constant δ_K of Φ is the smallest quantity such that

$$(1-\delta_K)\|x\|^2 \leq \|\Phi_T x\|^2 \leq (1+\delta_K)\|x\|^2 \tag{18.3}$$

for all the subset $T \subset N$, with $|T| \leq K$ and coefficient sequences $(x_j), j \in T$.

To check if a given matrix Φ satisfies the RIP is a difficult task. Fortunately, it can be shown that matrices with random entries drawn from certain probability distributions will have the RIP with high probability [1].

The signal x, which is K-sparse or compressible, can be recovered by solving the indeterminate system $y = \Phi x$, by (P1), from only $M \geq CK\log(N/K)$ samples, particularly with matrices Φ, with Gaussian entries [3].

The problem (P1) cannot be solved analytically, but can be reformulated as a linear programming problem when the data is real, and as a second order cone problem when the data is complex [4, 5]. In the complex case, $\|x\|_1$ is neither a linear nor a quadratic function of the real and imaginary components, and cannot be reformulated as one: $\| x \|_1 = \sum \sqrt{\Re(x_i)^2 + \Im(x_i)^2}$ In this case, the problem can be reformulated into *second order cone programming*, (SOCP), and solved with algorithms implemented in the framework of Interior Point Methods, for example using the CVX algorithm [6].

If a signal x can be written as a linear combination of K sinusoids, the signal presents few non-zero spectral lines in the classical Fourier Transform sense, that is, it is K-sparse in the frequency domain. However, in practical applications, because we use finite N-length signals, the signal is sparse only if the frequencies are multiples of the fundamental frequency $\frac{2\pi}{N}$. Leakage limits the success of the traditional CS algorithms. Here, we propose an iterative algorithm which find a first approximate location of a sinusoid and then refine the sampling in frequency around the neighborhood of this sinusoid up to a required precision. If several sinusoids have to be found, the procedure iterate as many times as needed this locale refinement.

The idea comes from the dual problem of fractional time-delay estimation, as studied in the work of Fuchs and Deylon [7]: the true value of the frequency can be obtained by BP between frequencies values having the higher values.

18.2 Spectral Estimation with Compressive Sensing

Consider $\Psi \in \mathbb{C}^{N\times N}$ as the inverse of the DFT matrix. Then, if x is a time domain discrete signal with length N, the DFT of x will be $s = \Psi^{-1}x$. If x is observed using random measurements, we have $y = \Phi x$, and we can write the problem (P1), from the Eq. (18.2):

$$\min \|s\|_1 : y = \Phi\Psi s = \Theta s, \tag{18.4}$$

The CS theory ensures that a signal that is sparse or compressible in the basis Ψ can be reconstructed from $M = O(K\log(\frac{N}{K}))$ linear projections onto a basis Φ that is incoherent with the first, solving the problem (P1) using the Eq. (18.4), [8, 9]. Random matrices are largely incoherent with any fixed basis [10].

If x contains only sinusoids with frequencies multiples of $\frac{2\pi}{N}$ rad, then s will be a sparse signal. Otherwise, s will be not sparse. If we apply the CS to solve this problem, the recovered signal s will not be sparse, as we can see in the example depicted in Fig. 18.1. The error is 0.5986 and even if we increase the number of measurements, the error remains large, having a value of 0.4368 for $M = 80$ and a value of 0.2892 for $M = 150$. Since the signal is not sparse we will need more measurements to get a better result.

This comes from the fact that no column vector in the matrix Ψ has a frequency matching one of the frequencies present in the signal. The first idea is to expand the matrix Ψ, so that each frequency present in the signal is represented by a column. We would have a redundant frame instead of the orthogonal basis of the DFT. By increasing the frame size, signals with frequencies that are not multiples of the fundamental frequency of the DFT become more compressible, resulting

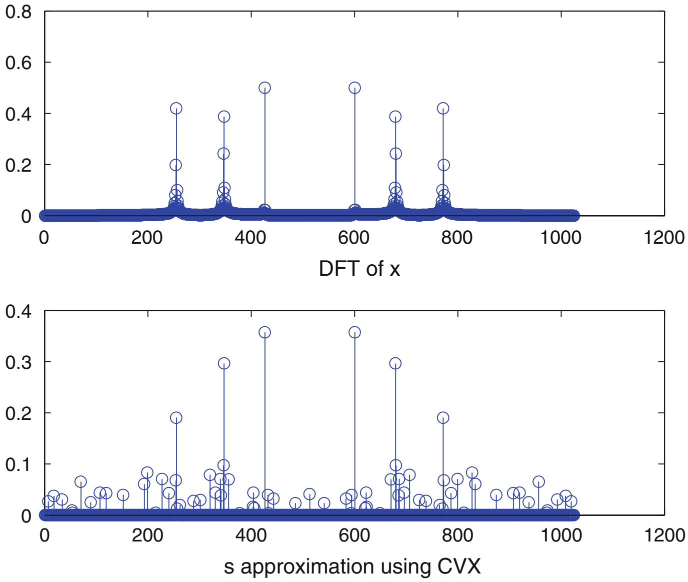


Fig. 18.1 Spectrum of a signal with length N=1,024 composed of three sinusoids that are not multiple of the fundamental frequency of the DFT, using the DFT and the CS using $M = 50$ measurements

into a recovery performance improvement, but in return, the frame becomes increasingly coherent, which leads to a decrease in performance recovery. So the idea was to add only a small number of columns. If we know between which columns of the Ψ matrix, are the frequencies that are not multiples of the fundamental frequency, we can add columns to the matrix Ψ, only in that interval, but in CS we only have access to the signal y.

18.2.1 Problem of Fractional Time-Delay Estimation

The dual problem of the fractional frequency spectral estimate is the fractional time-delay estimation. Fuchs and Deylon, in [11], presented an analytical expression of the minimal ℓ_1-norm interpolation function which is independent of the signal, to solve the problem to get an estimate of the delay τ, having a bandlimited signal $x(t)$, with a maximal sampling period $h = 1$, which is given by $y(t) = x(t - \tau)$. One possibility is to seek the values s_n in

$$y(t) = x(t - \tau) = \sum_n x(t - nh)s_n.$$

An estimation of the delay τ is determined from the maximum location of the interpolating function which is given by

$$\psi(t) = \sum_{k \geq 0} \beta_k \frac{\phi(|t| - k)}{|t|}, \quad |t| \in [k, k + \frac{1}{l}],$$

with

$$\phi(x) = \frac{1}{\Gamma(x)\Gamma(\frac{1}{l} - x)}, \quad x \in [0, \frac{1}{l}],$$

$$\beta_k = (-1)^k \frac{\Gamma(k + \frac{1}{l})}{\Gamma(k+1)},$$

where Γ is the standard gamma function. This reconstruction function is very localised and as the oversampling factor, l, increases more localised it will be, unlike what happens with the sinc function, which keeps the width of the main lobe, as one can see in Fig. 18.2.

Since the minimal ℓ_1- norm reconstruction function is quite localised, the s_n values can be obtained by solving the minimal ℓ_1-norm problem

$$\min \|s_n\|_1 : y(t) = \sum_n x(t - nh)s_n, \quad h < 1,$$

which is the BP.

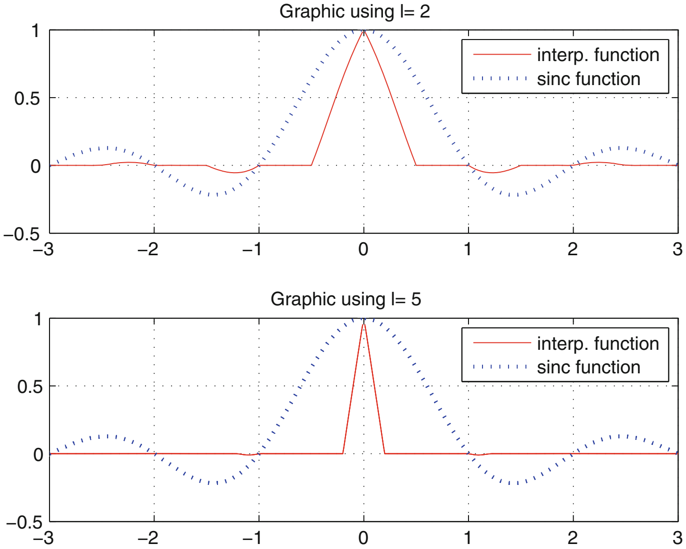


Fig. 18.2 ℓ_1-norm interpolating function with oversampling factors $l = 2$ and $l = 5$, compared with the sinc function

The problem we are dealing with is the dual of the problem studied by these authors. If we have a signal with a frequency f_i, which is a multiple of the fundamental frequency of the DFT, we know that the DFT of the signal has a maximum in the position of this frequency.

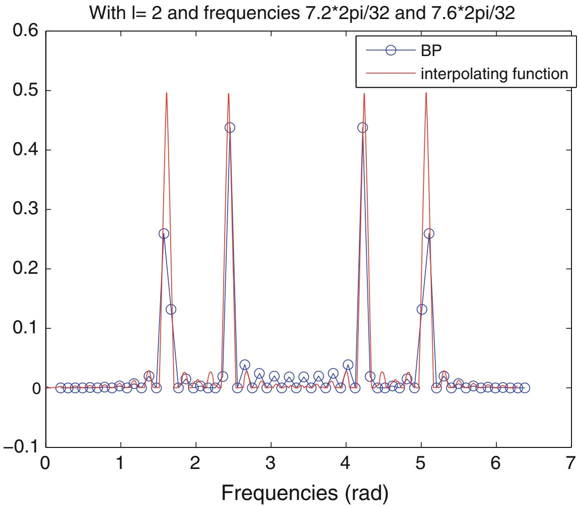
Looking to the frequency of the signal as the dual of the delay, the interpolating function will have a maximum exactly in the same place, independently of the considered l value. If we have a signal with a frequency f_i, which is not multiple of the DFT fundamental frequency, the signal is not sparse, therefore there are no maximums. However, the interpolating function will have a maximum in the position of the frequency, regardless of the amount of l which is considered. If the signal has two frequencies that are not multiples of the fundamental frequency, the interpolating function has two maximums, both between the values of the frequencies with higher values obtained by BP. See the example depicted in Fig. 18.3.

Thus, one possible solution to our problem of knowing where the frequencies are, is to apply the BP. Each of the frequencies in question, will be between the position of the two frequencies multiple of the fundamental frequency, where BP obtains maximum values.

If the frequencies are very close, a greater value of l must be used in order to discover them. See Fig. 18.4.

The proposed algorithm starts by finding the first interval where BP obtains maximum values. After that, we add columns among those corresponding to the endpoints of the interval. Then, we choose the column position where is the maximum value between the added columns, which is an estimation for the position of the desired frequency. Therefore, to determine the approximated value of another frequency, we expand the original matrix by adding that column. Then, by applying again the BP we find another interval and we repeat the same procedure.

Fig. 18.3 Minimal l_1 norm using BP and using the l_1 interpolating function



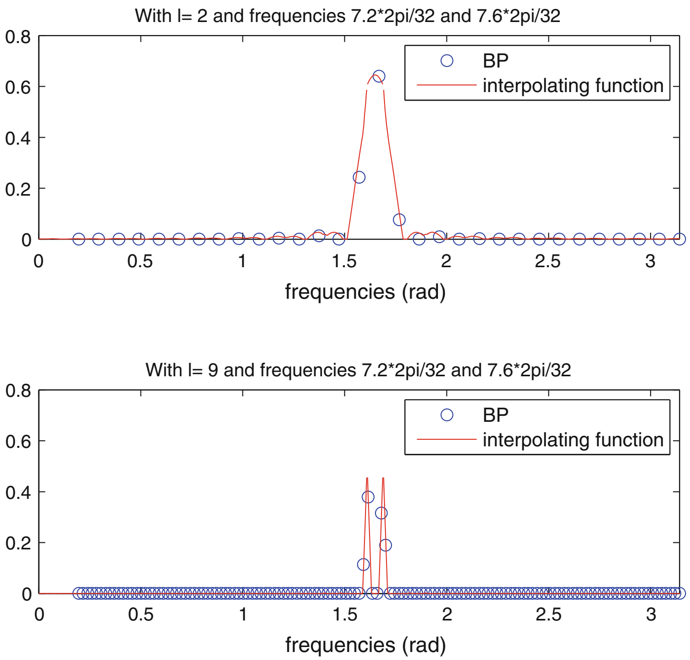


Fig. 18.4 Minimal l_1 norm using BP and using the interpolating function with frequencies very close, using two values of l

This algorithm can be easily extended for more frequencies as shown in the next section.

18.2.2 Proposed algorithm

We begin by calculating $\hat{s}$, which is the approximate value of s, solving the minimisation ℓ_1-norm problem:

$$\ell_1 : \min \|s\|_1 : y = \Theta s. \tag{18.5}$$

Then:

1. We will calculate the argmax of $\hat{s}$, s_{max};
2. The interval $[s_{max} - 1, s_{max}]$ or $[s_{max}, s_{max} + 1]$ is chosen as the image nearest to s_{max}. Let's call this interval [a,b];
3. We will add columns between the two extremes that correspond to the interval considered in the previous point:

```
$I := 0$
 while ($I <$ Nmaxpoint and $\varepsilon >$ error threshold)
```

a. $I := I + 1$
b. We will consider matrix Ψ_1, adding I columns to Ψ. The I frequencies of columns to add are given by: $(a + (1 : I)/(I + 1)) - 1$.
c. We will calculate the $\hat{s}$ values solving the problem 18.5 and considering the matrix Ψ_1;
d. We will calculate the argmax of $\hat{s}$, only in the interval $[a, b]$, which contain the I added columns;
e. We will consider a new matrix, Ψ_2, from Ψ, where in the interval $[a, b]$ is added the column which corresponds to the argmax of the values obtained in the previous point;
f. We will calculate the $\hat{s}$ values, using the matrix Ψ_2;
g. We compute the value of ε

4. $\Psi = \Psi_2$
5. We will repeat the steps from 1. to 4. as many times as the sparsity of the signal.

In the end we calculate the value $\hat{x} = \Psi\hat{s}$.

In this algorithm, we use the standard error, given by erro $= \frac{\|x-\hat{x}\|}{\|x\|}$. The stopping criterion in the reconstruction of the approximate value for each frequency, ε, i.e, the criteria used to stop adding columns in the range $[a, b]$ is given by the difference between the errors obtained in two consecutive iterations. In each iteration the error is given by the sum of absolute values of $\hat{s}$ excluding the K higher values, with K the value of sparsity. If in the interval $[a, b]$, on the step 3f., we add the column corresponding to the frequency of the sinusoid, this error is very small [10].

18.3 Experimental Results

In our experiments we use signals of length $N = 1{,}024$ samples and all the signals contain real-value sinusoids, with random frequencies. The amplitudes of each frequency is 1 except in the experiment III.

18.3.1 Experiment I

In our first experiment, we apply the proposed algorithm to a signal containing three real-value sinusoids, $K = 6$, using $M = 100$ measurements. Therefore we can reconstruct the signal with an error of 0.0268, which had an initial error of 0.5275, see Fig. 18.5.

As shown in Fig. 18.6, the error decreases exponentially as we add columns in the interval, so we can initialise the number of adding columns, step 1. of the proposed algorithm, with a greater value than $I = 1$. In our experiments we initialised with $I = 650$.

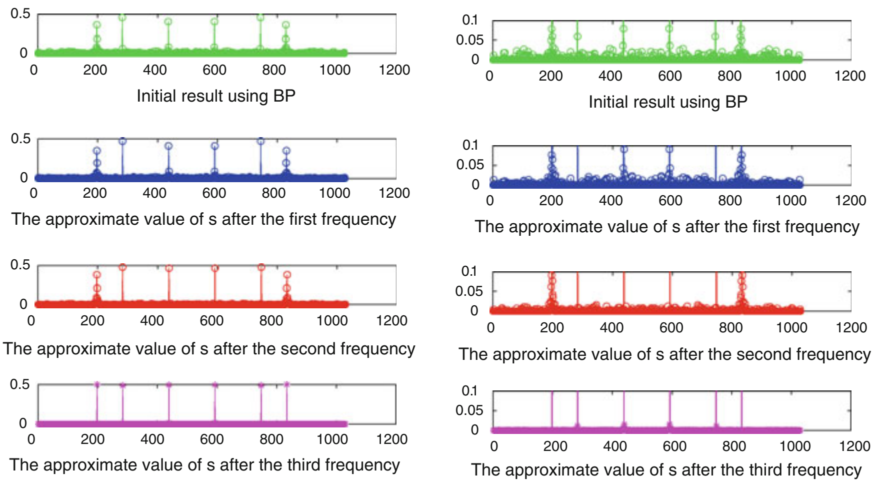


Fig. 18.5 The approximate value of s, first using CVX to solve the BP, and then with the proposed algorithm in the first, second and third frequencies

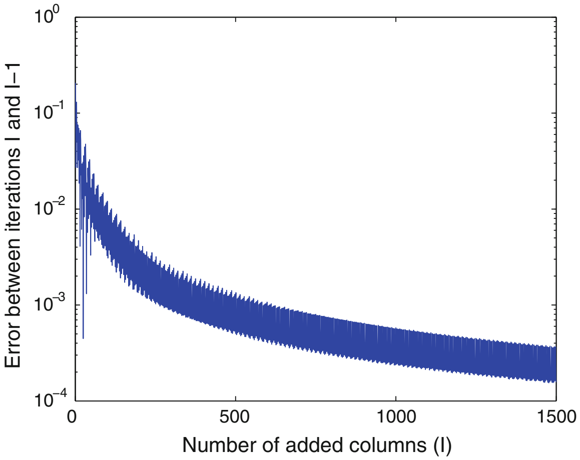
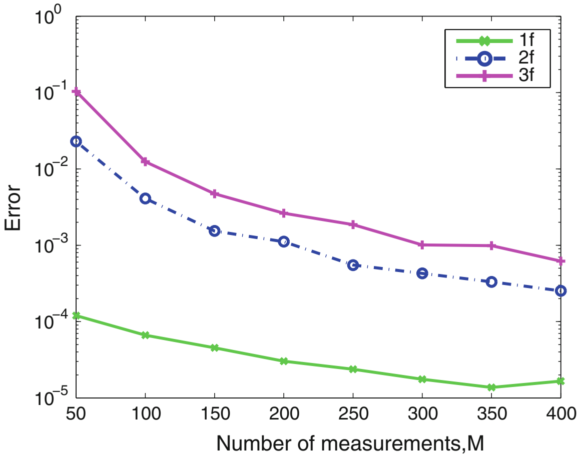


Fig. 18.6 The error in the first frequency in function of the number of added columns

18.3.2 Experiment II

Our second experiment compares the performance of the proposed algorithm for signals with one, two and three frequencies. We verify that the number of measurements we need for the same performance increases with the sparseness of the signal. See Fig. 18.7.

Fig. 18.7 Performance of CS signal recovery with the proposed algorithm for signals with one, two, three and four frequencies which correspond to $K = 2$, $K = 4$ and $K = 6$ respectively. All quantities are averaged over 400 independent trials



18.3.3 Experiment III

This experiment shows the behaviour of the proposed algorithm, when the amplitudes of the signal frequencies are different. Figure 18.8 presents the result of the signal recovery using our algorithm for a signal composed by three random frequencies with amplitudes 1, 0.01 and 0.05. The proposed algorithm performs better for different amplitudes, than Thresholding based algorithms, like Spectral Iterative Hard Thresholding (SIHT) proposed by M. Duarte and et al. in [12], since the Thresholding algorithms consider, in each iteration, only the K largest spectral components, removing the others. With this approach, the smallest frequencies can be discarded.

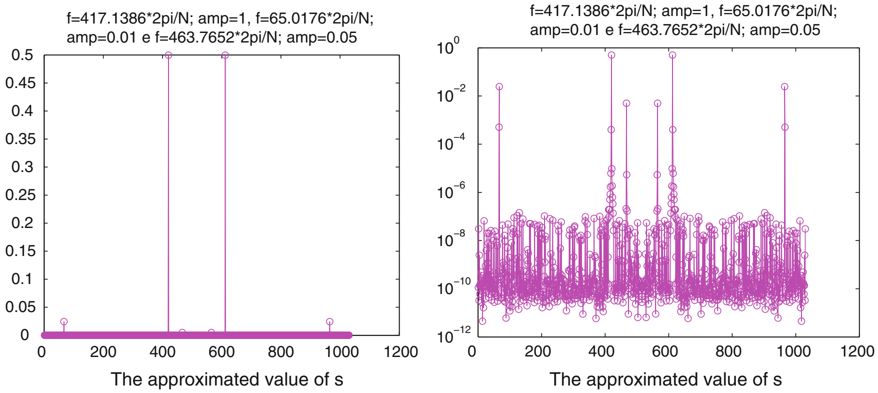


Fig. 18.8 The approximate value s, for a signal containing three real-value sinusoids with frequencies of amplitudes 1, 0.01 and 0.05

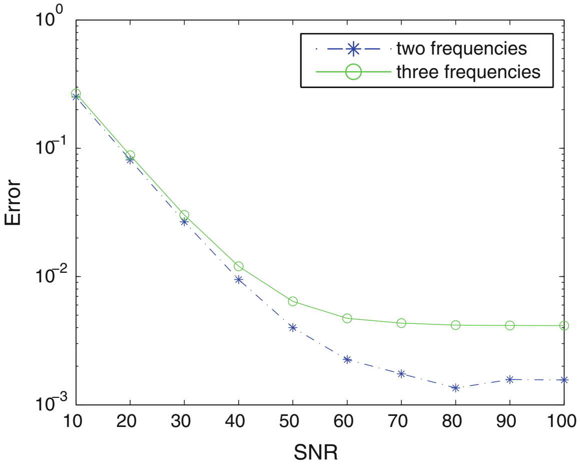


Fig. 18.9 Performance of signal recovery using the proposed algorithm for a signal composed by two and three different frequencies with 150 noisy measurements. All quantities are averaged over 100 independent trials

18.3.4 Experiment IV

Our fourth experiment tests the robustness of the proposed algorithm to additive noise in the measurements of a signal written as a linear combination of two and three sinusoids. The error was evaluated for ten signal to noise ratios (SNR) and the results are depicted in Fig. 18.9. As we can see, the proposed algorithm performs quite well.

18.3.5 Experiment V

This experiment compares the performance of the proposed algorithm with two of the algorithms proposed by M. Duarte and et al. in [3], which they called by Spectral CS (SCS), Spectral Iterative Hard Thresholding (SIHT) using a heuristic approximation and a Line Spectral Estimation (Root Music). See Fig. 18.10.

For the SIHT, the authors use an over-sampled DFT frame and a coherent-inhibiting structured signal model, that inhibits closely spaced sinusoids, and the classical sinusoid parameter estimation algorithm, periodogram. In our algorithm we do not need to impose a model based, to inhibit the coherence of the frame, because our interpolating function is very localised.

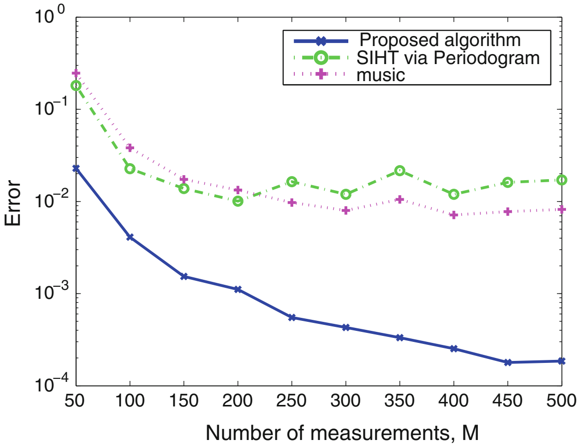


Fig. 18.10 Performance of signal recovery using the proposed algorithm, using the SIHT implemented via heuristic algorithm and using the Root Music algorithm. All quantities are averaged over 400 independent trials

18.3.6 Experiment VI

Our last experiment shows how the proposed algorithm behaves for two close frequencies and compares its performance with the performance of the SIHT and the Root Music algorithms. We have considered a fixed frequency, f_1 and a second frequency, $f_2 = f_1 + \delta$, where $\delta = [0.1 : 0.1 : 1,\ 1.25 : 0.25 : 5.5]$. As shown in Fig. 18.11, our algorithm presents a better performance than the others.

Note that, although the errors are small for delta values smaller than 1, the frequencies values can be very different from the correct ones, as we can see in Table 18.1

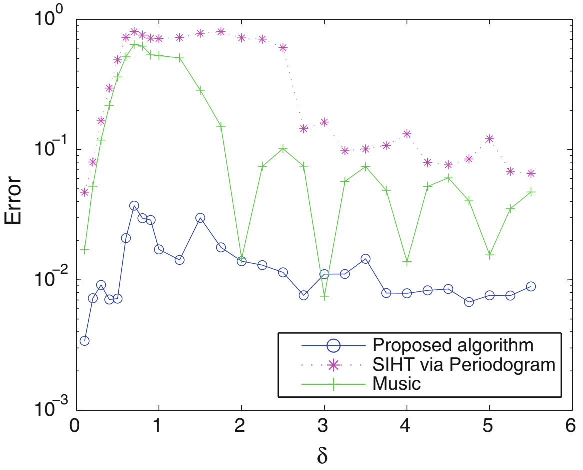


Fig. 18.11 Performance of signal recovery using the proposed algorithm, the SIHT implemented via heuristic algorithm and the Root Music algorithm, considering a signal with two frequencies where $f_2 = f_1 + \delta$. We have used 100 measurements. All quantities are averaged over 200 independent trials

Table 18.1 Recovered values obtained with the proposed algorithm, the Root Music algorithm and the siht algorithm, using the fixed frequency f_1 and $f_2 = f_1 + \delta$

	Frequencies		**Prop. Algorithm**		**MUSIC**		**SIHT**	
δ	f_1	f_2	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_1$	$\hat{f}_2$
0.1000	463.7655	463.8655	463.7638	464.8000	78.8620	463.8151	0.0000	0.0000
0.2000	463.7655	463.9655	463.7621	464.8000	52.4859	463.8594	460.4439	463.8723
0.3000	463.7655	464.0655	464.0734	464.0000	265.7705	463.9192	460.4508	463.9125
0.4000	463.7655	464.1655	463.8000	464.2148	67.0411	463.9667	22.8131	463.9436
0.5000	463.7655	464.2655	464.2000	463.7007	464.0346	469.9109	378.3570	463.8356
0.6000	463.7655	464.3655	463.9000	464.4437	0.0377	397.5788	459.2702	463.8318
0.7000	463.7655	464.4655	463.4965	464.5676	370.1165	464.2669	246.8810	463.3999
0.8000	463.7655	464.5655	464.6005	463.7246	212.0844	463.8000	129.2858	464.9987
0.9000	463.7655	464.6655	464.6176	463.7699	408.0093	463.6348	255.1839	463.5699
1.0000	463.7655	464.7655	463.5866	464.8085	406.0506	464.8486	199.8234	464.9885

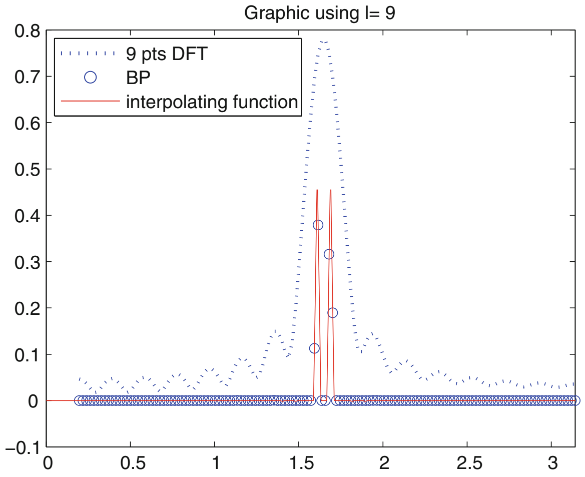


Fig. 18.12 Minimal ℓ_1-norm using BP and using the interpolating function with an over sampling factor of $l = 9$. The dotted curve is the minimal ℓ_2-norm interpolating function: the sinc function

The results of the proposed algorithm were as it was expected, since the minimal ℓ_1-norm interpolation function is very localised, unlike the minimal ℓ_2-norm interpolation function - the sinc function, as one can see in Fig. 18.12.

18.4 Conclusion

We have developed a new algorithm to estimate the spectral components in the case of sparse finite-length signals. The algorithm uses a redundant frame, transforming the DFT basis into a frame with a larger number of vectors, by inserting columns between some of the initial ones. The frame has a maximum of $N + K$ vectors, with K the sparsity of the signal.

From the results can be seen that the proposed algorithm can recover the sparse signals with an error smaller than 0.001, even for a signal with $K = 6$.

Furthermore, it presents a good performance in the presence of noise. In addition to this, it can deal with signals where the frequency amplitudes are very different, overcoming other algorithms in this field. Moreover, the proposed algorithm performs better than others that we have compared for the same signal while using the same number of measurements.

References

1. Baraniuk R, Davenport M, DeVore R, Wakin M (2008) A simple proof of the restricted isometry property for random matrices. Constructive Approximation, 28:(3), 253–263
2. Donoho DL, Huo X (2001) Uncertainty principles and ideal atomic decomposition. IEEE Trans Inf Theory 47:2845–2862
3. Tropp JA, Gilbert AC (2007) Signal recovery from random measurements via orthogonal matching pursuit. IEEE Trans Inf Theory 53:4655–4666
4. Malioutov DM (2003) A sparse signal reconstruction perspective for source localization with sensor arrays. MIT
5. Winter S, Sawada H Makino S (2005) On real and complex valued $l1$-norm minimization for overcomplete blind source separation. 86–89
6. Boyd S, Grant M (2011) Matlab software for disciplined convex programming, v. 1.21.
7. Fuchs JJ, Delyon B (2000) Minimal L1-norm reconstruction function for oversampled signals: applications to time-delay estimation. IEEE Trans Inf Theory 46:1666–1673
8. Candés E, Romberg J, Tao T (2006) Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans Inf Theory 52:489–509
9. Donoho DL (2006) Compressed sensing. IEEE Trans Inf Theory 52:1289–1306
10. Candés E, Wakin MB (2008) An introduction to compressive sampling. IEEE Signal Process Mag 25–2:21–30
11. Duarte I, Vieira J, Ferreira P, Albuquerque, D (2012) High resolution spectral estimation using BP via compressive sensing. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science, WCECS 2012, 24–26 Oct 2012, San Francisco, USA, 699–704
12. Baraniuk RG, Duarte M F (2010) Spectral compressive sensing submitted to, IEEE transactions on signal processing
13. Candés E, Romberg J, Tao T (2006) Stable signal recovery from incomplete and inaccurate measurements. Commun Pure Appl Math 59–8:1207–1223
14. Donoho DL, Bruckstein AM, Elad M (2009) From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev 51:34–81

Chapter 19
Stereoscopic 3D Adjustment Under Specific Capture Conditions

Tzung-Han Lin

Abstract This chapter addresses a practical method for adjusting stereo camera parameters under specific capture conditions. We use computer generated models and motion capture data to simulate various action scenarios. Our goal is to keep an appropriate 3D effect for the interested character who has a specific motion. This method analyses the parallax distribution for every frame, and the parallax distribution parallax distribution is used to adjust the parameters of a virtual stereo camera for quality depth perception.

Keywords Depth perception · Dynamic 3D adjustment · Motion capture · Parallax distribution · Stereo camera · Stereoscopic images · Visual comfort

19.1 Introduction

Production of quality 3D content is a difficult art that requires a variety of technical, psychological, and creative skills. It has to consider human perception, display capabilities and viewing conditions. Most problems of stereo images may occur due to improper operations of stereo cameras. To obtain quality 3D content, a smart control method for the stereo camera becomes important [1]. Recently, stereo camera has become a programmable device, and it is capable of running a specific script for adapting the stereo camera parameters [2]. The programmable stereo device can simultaneously capture scenes and provide disparity information by analyzing stereo pairs. Then, stereo camera parameters are adjusted for changing 3D effect in either automatic ways or manual operations. However, the production of quality 3D content is still costly. Capturing dynamic stereo scenes is

T.-H. Lin (✉)
National Taiwan University of Science and Technology,
No. 43, Sec. 4, Keelung Rd, Taipei 10607, Taiwan
e-mail: thl@mail.ntust.edu.tw

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_19,

much difficult, since it involves temporal change. Sometime the analysis of one static image is not applicable for a dynamic scene. Therefore, the parameter script generator for various action scenarios is getting important.

However, there are many physical constraints for stereo cameras when people are taking 3D films [3]. Obviously, the position of the stereo camera and the capturing direction are usually controlled by stereographers, and these two parameters are the native constraints for forming stereoscopic images. It is still difficult for stereographers to handle all the rest parameters of the stereo cameras when they are taking 3D films. Besides, the stereo camera parameters are dependent. Based on the need, we develop a tool to analyze parallax distributions. The parallax distribution provides the key feature for preventing viewing discomfort. Finally, the parameter scripts are generated and used again for reproducing the same 3D effect in similar situations.

19.2 Background

The visual comfort of stereoscopic images is the most critical problem in stereoscopic researches [4]. It regards the conflict between the accommodation and the convergence of human eyes. It usually refers to the subjective sensation. However, there is no standard methodology for the measurement of visual comfort in stereoscopic images. From the recommendation of the International Telecommunications Union (ITU), it only considers the picture and depth quality on subjective methods [5]. Despite that, the limits of visual comfort are suggested as specific disparity ranges and specific values in various viewing conditions [6]. The range of the disparity becomes an important paradigm for producing stereoscopic images. Lambooij et al. [7] found that not only the magnitude but also the distribution of disparities affects visual comfort. visual comfort can be modeled as a combined effect of screen disparity range with to a lesser extent screen disparity offset, changing screen disparity and lateral motion of which the specific contributions depend on the activity of the scene.

Jones et al. [8] proposed a controllable perceived depth method for generating stereoscopic images from parallel cameras. Their method transforms the scene depth to a specific perceived depth. The depth distortion is also avoided in head tracked displays; Sun and Holliman [9] used a subjective human-based experiment to evaluate different stereoscopic algorithms. The result shows the practical 3D viewing volume differs between individual displays. Generally speaking, the comfortable 3D viewing ranges are expanded in viewing dynamic stereoscopic images in comparison to static stereoscopic images.

Quintus and Halle [10] developed a composition tool for creating comfortable stereoscopic images from static 3D digital models. In their work, camera position, camera view angle, projection plane, viewing distance and interaxial distance are considered for adapting a comfortable stereoscopic image. Their tool also provides realistic visualization to assist doctors or engineers having better depth judgment in their professional operations.

Pockett and Salmimaa [11] proposed a quality improvement method for user created stereoscopic content. Their method optimizes the disparity range under a parallel camera configuration. As a consequence, a specific 3D effect is guaranteed for mobile device displays.

We have experienced the 3D effects of motion capture data under various parameters. The simulation environment is the puppet with motion data to simulate real capture results. Then, the simulation system is developed for analysis of the parallax distributions

of stereoscopic images. In our work, the parallax distribution is the main factor for quality 3D content. With the simulation system, desired camera parameters can be used for the design of real stereo cameras.

19.3 Methods

19.3.1 Virtual Stereo Camera

In our configuration, the virtual stereo camera system consists of two identical cameras. Their field of view (FOV), convergence angle and interaxial distance are the independent parameters used for adapting the 3D effect as in Fig. 19.1. Various 3D effects are induced by assigning different parameters, but 3D effects should be acceptable for visual comfort. Although, our method adapts the 3D effect by adjusting stereo parameters, not all parameters produce good 3D content, especially for motion scenes. In most conditions, the stereo camera system has physical limitations. For example, the interaxial distance can't be very small. The convergence angle is very unlikely to be a negative value, since the divergent stereo pairs are very difficult to be fused by human brain.

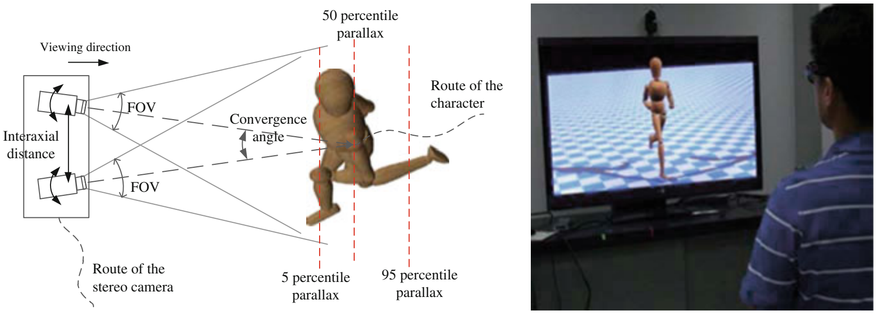


Fig. 19.1 Schematics of the virtual stereo camera configuration and the 3D scene on the display. In *left figure*, three parameters in the camera-rig are used for adjusting the 3D effect automatically. The stereoscopic image is synthesized and its parallax distribution is changed to fit the given constraints and specific thresholds shown in *right figure*

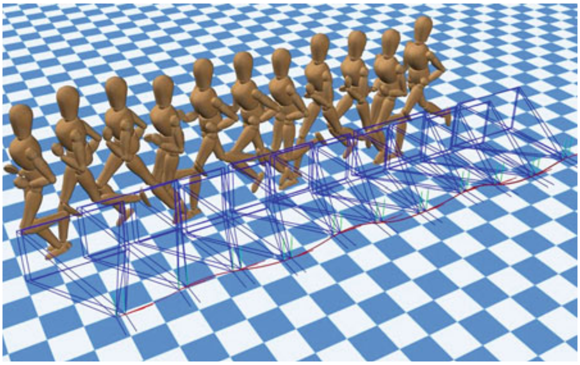


Fig. 19.2 Schematic of the native constraints in photography: the stereoscopic camera is constrained on a spline path for capturing a running character. Our method dynamically adjusts parameters according to its instant parallax distribution

Figure 19.2 illustrates one of the application scenarios. In this scenario, the stereo camera is supposed to follow the path of a running man. Furthermore, we expect that the 3D effect in every frame is steady. That means their parallax distributions are similar. Base on this requirement, we create a simulation system by openGL to evaluate the 3D stereoscopic effect. The virtual stereo camera and one character with a specific motion are defined. The motion data of the character are from CMU motion capture database [12]. We assume both views of the stereo camera have the same perspective effect without lens distortion. We do not concern their physical discrepancies, such as color or brightness. All frames are synchronized. And the up vector of the stereo camera should be carefully handled. Our method considers five parameters. They are stereo convergence angle, interaxial distance, field of view (FOV), camera position and viewing direction. In our simulation system, all parameters are either controlled by the stereographer or constrained by our method for rendering different 3D effects.

19.3.2 Parallax Analysis

Parallax distribution is an important feature for assessment of visual comfort in stereoscopic images [13, 14]. The visual comfort regards not only parallax magnitudes but also parallax dispersions. In dynamic stereoscopic scenes, the changing disparity will affect the visual comfort obviously [15]. However, the stereoscopic image needs stereo matching algorithms to have disparity maps. For convenience, we use 3D computer graphics simulation for generating disparity maps. The flowchart is shown in Fig. 19.3. Since the simulated scenes are computer generated, it is easy to acquire binocular depth buffer for analysis in real-time. The parallax data from one stereoscopic frame are converted into a distribution histogram, and only the pixels of the character are considered. In the histogram, we calculate the distribution centroid as the mean parallax. The near limit and far limit of the visual comfort are considered as the 5 percentile and 95 percentile in the parallax cumulative histogram, respectively. In other words, we consider 5

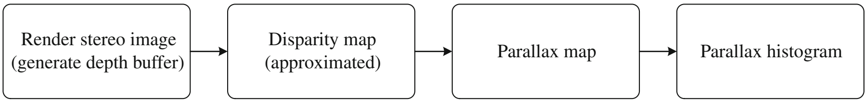
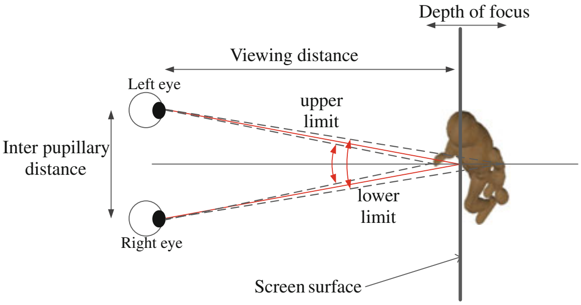


Fig. 19.3 Procedure for obtaining parallax behaviors

Fig. 19.4 Schematic of upper and lower limits of parallax when watching 3D videos



percentile and 95 percentile parallaxes as the upper and lower limits, respectively. These values are corresponding to the expected limits of depth of focus (DOF) as shown in Fig. 19.4.

19.3.3 Parallax Adjustment

In order to preserve the 3D effect as the same with initial conditions, all stereo parameters are calculated from the stereo parameters of the first frame. Our method alternatively adjusts the convergence angle and the interaxial distance for changing parallax distribution under the given constraints. Figure 19.5 shows the flowchart of our method. Initially, a character motion and stereo camera positions are given. Then, our method adjusts the viewing direction and parameters according to the current status. Since we can't expect where the stereo camera is, these adjustments for stereo camera parameters highly depend on their initial conditions. However, the parallax distribution is calculated from the stereo image. The way we change its distribution is to adjust the interaxial distance and the convergence angle. A larger interaxial distance induces a more hyper stereo effect, and it depends on how far the character is. Our goal is to keep the mean parallax zero in the first iteration. And then, the adjustment of the convergence angle will enlarge or suppress the parallax range. To avoid visual discomfort, we usually set a threshold for the parallax range. For an exaggerating 3D effect, the parallax range is often larger than 1°.

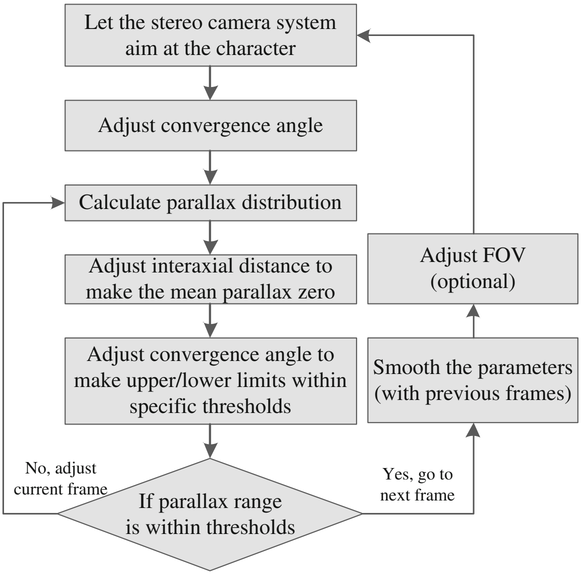

Let the stereo camera system aim at the character
Adjust convergence angle
Calculate parallax distribution
Adjust interaxial distance to make the mean parallax zero
Adjust convergence angle to make upper/lower limits within specific thresholds
If parallax range is within thresholds
No, adjust current frame
Yes, go to next frame
Smooth the parameters (with previous frames)
Adjust FOV (optional)

Fig. 19.5 Flowchart of the proposed method. Our method alternatively adjusts parameters under the given constraints. The FOV is the optional parameter to change the perspective effect

19.3.4 Limitation of Method

However, our method has some limitations. In many films, the character in specific actions always attracts the audiences. So, we assume that audiences frequently focus their looks at the single character in the central area of the 3D display. To put the character on the center area of the 3D display, the viewing direction is restricted to pass through the locus of the character's bounding box. However, it may make the video jagging. We apply a smooth operation on all the parameters to suppress the jagging phenomenon as shown in Fig. 19.6. It can be done by Kalman filter algorithm, as well. Our method only considers parallax distribution of the character. The result will depend on the distance between the stereo camera and

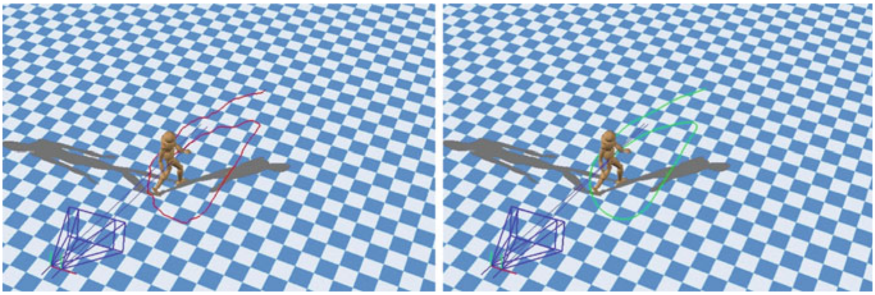


Fig. 19.6 Two examples show the difference between real and simulated cases. A real route of the bounding box is illustrated in *left figure*, and the simulated route is smoothed in *right figure*

the character. The limitation of our method is that the motion of character and the route of the camera should be continuous.

19.4 Result and Discussion

Our experiment device is a 27″ polarized stereoscopic display. The developed program is built for generating stereoscopic images from motion capture data. The disparities of stereoscopic images are calculated from the depth buffers of left view images only. Then, the parallaxes are converted from the disparities according to the viewing condition. We assume the viewing distance is 90 cm and the pupillary distance is 6.5 cm. In our experiment, the virtual stereo camera generates 60 stereoscopic images per second, and the resolution is 1,920 by 1,080. The example in Fig. 19.7 shows the parallax distribution of a kicking character before adjustment. In its parallax distribution, the mean value is used for changing the interaxial distance at first iteration, and then the parallax range is used for adjusting convergence angle, alternatively. The near limit and far limit are at the 5 percentile and 95 percentile of cumulative probability density, respectively. Then, the parallax distribution of one stereoscopic image is obtained.

In Figs. 19.8 and 19.9, two different initial conditions with the same motion are shown. In Fig. 19.8, a small interaxial distance is given for a weak 3D effect initially. To keep the approximate parallax range compared with the first frame, two parameters are adjusted according to parallax distributions. Due to the geometric relation, we simultaneously adjust these two parameters. Consequently, their variations are reduced. If one of the two parameters has a physical limit, the other parameter still remains 1° of freedom for adjustment. Our method keeps the

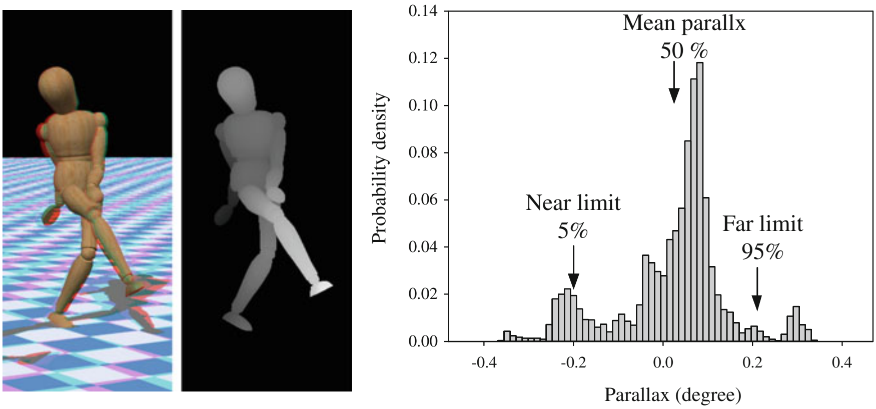


Fig. 19.7 Parallax analysis. A cropped snapshot of the character is shown in anaglyph (*left*). Only the depth buffer of the character is considered for analysis (*middle*). Its parallax distribution is converted from the depth buffer according to the viewing condition in *right*

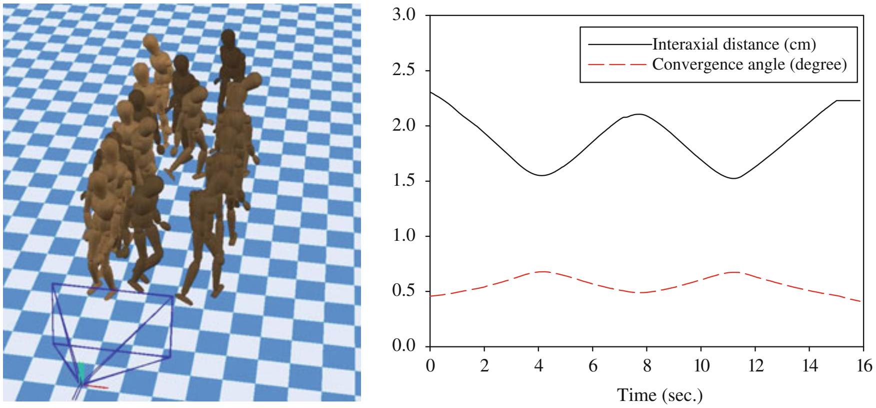


Fig. 19.8 A walk motion and a fixed camera position are shown in the *left figure*. Initially, the interaxial distance is small. A man walked twice forward and then backward. The brighter shaded model is at the later position. All generated parameters are plotted in the *right figure*

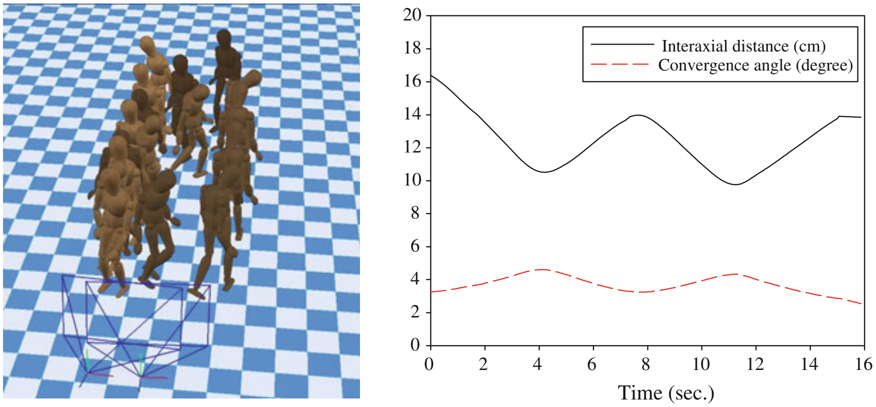


Fig. 19.9 A large interaxial distance is given initially shown in *left figure*. The parameter output is plotted in *right figure*)

zero-parallax plane on the centroid of the parallax distribution as possible. Nevertheless, changing the interaxial distance is not the only way for adjusting the value of the mean parallax. For traditional parallel stereo cameras, shifting both images inward or outward is often used for reproducing its parallax distribution. This skill will affect the position of zero parallax only. And it may induce blank pixels on the left and right borders. For the case of a large initial interaxial distance, the similar result is shown in the right part of Fig. 19.9.

Figure 19.10 shows the character is walking to the stereo camera whose position is fixed. The FOV parameter is independent and optional. It is used for changing the perspective effect. In this figure, FOV is defined as the function of the

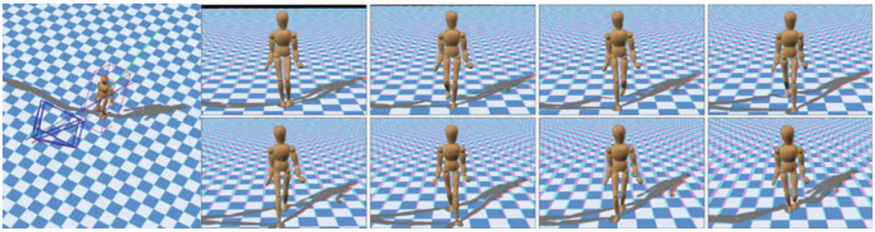


Fig. 19.10 The character walks forward and the stereo camera position is fixed (*left figure*). The images with the same parallax range are generated (*rest figures*)

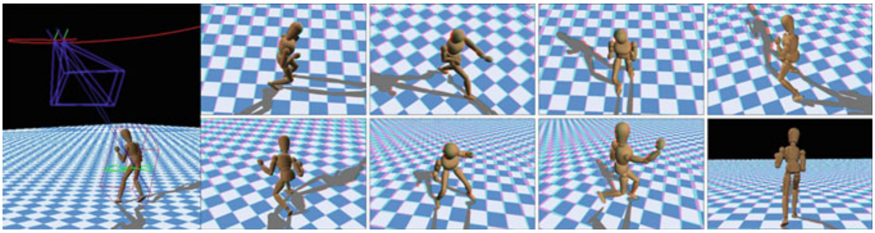


Fig. 19.11 The character is boxing and the camera moves along a top-down spiral path (*left figure*). The selected frames are shown (*rest figures*)

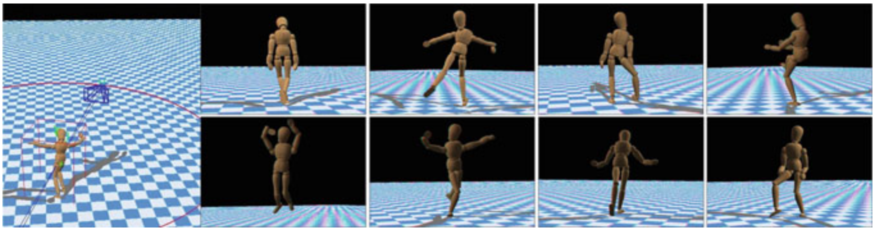


Fig. 19.12 The character is dancing and the camera moves along a circular path (*left figure*). The rendering results are illustrated (*rest figures*)

character's size, and that will make its size consistent. When the character comes close to the camera, its parallax range is almost the same compared with the initial value.

The camera position is often handled by the stereographer, either along a predefined path or on arbitrary routes. Since the stereo camera has 6° of freedom in Euclidean space, we keep its roll angle constant to avoid unnatural images. In a real case, the pose of the camera can be readily detected by a gyro for compensation. This is a basic requirement in our test system. In Fig. 19.11, we simulate 3D effects with the stereo camera on a spiral path. The camera moves from top to bottom and always aims at the boxing character. When the camera moves, the camera's pose will be corrected. Then, the stereo camera parameters are updated.

Another example shown in Fig. 19.12 is a dancing character with a stereo camera on a circular path. In this example, the viewing direction is calculated to focus on the smooth path of this character.

Although our method keeps the mean parallax zero and makes the parallax range controllable, the 3D effect is sometime subjective. It is worth to generate parameter scripts for various kinds of scenarios, since the simulation data provide different experiences for capturing common motions as stereoscopic films. The same scenario frequently happens in sports broadcasting and action films. Recently, the commercial camera with intelligent functions for assisting photographers has become a trend. Generating specific 3D effects in photography may become routine. Besides, our method does not consider the background parallax. This is because our test conditions are dynamic and the interested object is the character motion.

19.5 Conclusion and Future Work

We carried out a simulation system for capturing stereoscopic motion characters under a constrained camera path and initial parallax conditions. The output can be parameter scripts for desired 3D effects. With regard of future issues for study, additional subjective experiments for favor 3D effects should be conducted.

Acknowledgments This work was supported in part by the National Science Council of Taiwan under Grant NSC100-2218-E-011-028.

References

1. Lin TH (2012) Controlling depth perception of stereoscopic images under given constraints. Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct 2012, San Francisco, USA, pp 630–633
2. Heinzle S et al (2011) Computational stereo camera system with programmable control loop. ACM Trans Graphics 30(4):1–10
3. Mendiburu B (2009) 3D movie making stereoscopic digital cinema from script to screen. Focal Press, Oxford Ch. 5
4. Hoffman DM, Girshick AR, Banks MS (2008) Vergence– accommodation conflicts hinder visual performance and cause visual fatigue. J Vis 8:1–30
5. Subjective assessment of stereoscopic television pictures (2000) ITU-R Recommendation BT.1438
6. Tam W, Speranza F, Yano S (2011) Stereoscopic 3D-TV: visual comfort. IEEE Trans Broadcast 57(2):335–346
7. Lambooij M, Fortuin M, IJsselsteijn W, Heynderickx I (2010) Visual discomfort associated with 3D displays. In: International workshop on video processing and quality metrics for consumer electronics, Arizona

8. Jones G, Lee D, Holliman N, Ezra D (2001) Controlling perceived depth in stereoscopic images. In: Proceedings of SPIE, 2001, vol 4297, pp 42–53
9. Sun G, Holliman N (2009) Evaluating methods for controlling depth perception in stereoscopic cinematography. In: Proceedings of SPIE, 2009, vol 7237, pp 1–12
10. Quintus K, Halle M (2008) A composition tool for creating comfortable stereoscopic images. In: Proceedings of SPIE, 2008, vol 6803, pp 1–10
11. Pockett LD, Salmimaa MP (2008) Methods for improving the quality of user created stereoscopic content. In: Proceedings of SPIE, 2008, vol 6803, pp 1–11
12. CMU motion capture database, http://mocap.cs.cmu.edu/
13. Nojiri Y (2003) Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV. In: Proceedings of SPIE, 2003, vol 5006, pp 195–205
14. Nojiri Y, Yamanoue H, Ide S, Yano S, Okana F (2006) Parallax distribution and visual comfort on stereoscopic HDTV. In: Proceedings of IBC, 2006, no 3, pp 373–380
15. Speranza F (2006) Effect of disparity and motion on visual comfort of stereoscopic images. In: Proceedings of SPIE, 2006, vol 6055, pp 1–10

Chapter 20
Cursive Handwritten Text Document Preprocessing Methodologies

Neeta Nain and Subhash Panwar

Abstract Handwritten text recognition offers a new way of improving the human computer interface and offer integrating computers better into human society. This chapter details the complete preprocessing stage in handwritten text document recognition. The input for this type of text document recognition system is a scanned image of handwritten text and the output is the normalized and segmented text lines. We have explained simple and efficient general sub steps of preprocessing namely binarization, line separation, skew normalization and slant correction of handwritten text. For binarization we have proposed a fast adaptive approach which gives good results. Line segmentation is done using bottom up grouping approach. We have further used a connectivity strength parameter for extraction of connected components (strokes) of same line from minimum spanning tree of given connected components. Quantitative analysis shows that this approach gives better results compared to others for line separation in the presence of touched or intermingled lines and also solving the Hill-and-Dale writing styles. Further for Skew normalization we use orthogonal projection approach which detects the exact skew angle without calculating the baseline separately. A variety of approaches have been explored for each step. We present a complete preprocessing suite for handling cursive handwritten text documents.

Keywords Binarization · Handwritten text recognition · Line segmentation · Normalization · Preprocessing · Skew

N. Nain · S. Panwar (✉)
Malaviya National Institute of Technology Jaipur, JLN Marg, Jaipur 302017, India
e-mail: panwar.subhash@gmail.com

N. Nain
e-mail: neetanain@yahoo.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_20,

20.1 Introduction

Now a days we are in a era of electronics document management system and it gives great benefit to society. The format understood by computer as word processor, computer aided design (CAD) and mark-up language are extensively used to process the data. These process may be storage, copy, editing and retrieval as on the documents. The Historical documents also need safe storage in the computerized format. Such formatted documents can be easily edited, copied on paper or may be distributed electronically across world wide networks. On the other hand, when documents are on paper, computer can not perform all such task. The manual data entry is required to change the document in computer understanding format but it is a very costing process and most of this cost is in human labour. When the process of data entry is automated, significant cost can be reduced. In all of the application, the major goal is to extract the information contained in the documents and store them in a computerized format for recognition. The handwritten document are also need recognition to store in computer understanding format. As shown in [1] the field of handwritten recognition can be divided into on-line and off-line recognition. In on-line recognition the writer is physically connected to a computer via a mouse, an electronic pen, or a touch sensitive device and his or her handwriting is recorded as a time dependent process. By contrast, in the off-line mode handwriting is captured by means of a scanner and becomes available in form of an image without any temporal information. Also the use of cameras for capturing handwriting is becoming increasingly popular. Because of the lack of temporal information, off-line recognition is considered the more difficult problem. we will focus our attention on off-line recognition. However, it has to be noted that there are close relations between the two modalities on-line and off-line recognition. Because in both off-line and online problem generally features vectors are used by various methods after the writing. Handwritten text recognition as shown in [2] generally having four basic steps: (1) Preprocessing (2) Segmentation (3) Feature Extraction (4) Classification. In the preprocessing step we convert Gray level image into bi level image and this step is known as binarization. After the binarization, image is often filtered to remove noise using suitable filters.

The Segmentation of a document image into its basic entities, namely, text lines and words, is considered as a non trivial problem to solve in the field of handwritten document recognition. The line of text document is segmented using projection profile, Hough transformation and smearing methods. Segmentation step is used in two phase. first phase of segmentation is done after binarization as ling segmentation and the second phase of segmentation is done after completion of preprocessing step. Here we will discuss a new approach for line segmentation which handle touching line problem also. After the line segmentation the Connected component are used for word and character segmentation. There are grouping methods and graph based methods for finding the connection between the several components of text documents. Generally the distance is calculated in between the components and we design the Minimum spanning tree using the component as vertices of graphs and the calculated distance of components are used as edges of graph and a predefined minimum distance are used for differentiate between in the inter word and intra word distance.

To reducing the dimension of segmented image of characters the feature vector is designed. Two kinds of features can be extracted from the segmented images. The features based on its statistical information as projection profile are called statistical features as derived from statistical distribution of points, and Structural features are based on topological and geometrical properties of the character, such as maxima and minima, size and shape, branch points, strokes and their directions, inflection between two points, horizontal curves at top or bottom, etc. The features of each test characters are extracted and these feature vectors to be used as input to classifier and then feature vector of each segmented text character is classified. There is various classification approaches are available for text recognition as K-nearest classifier, Bayes classifier, Polynomial classifiers, neural network and support vector machine.

20.2 Preprocessing

Preprocessing aims to produce clean document images so that recognition is done accurately. It is done prior to the application of segmentation and feature extraction algorithms. The raw data is subjected to a number of preliminary processing steps to make it usable in the descriptive stages of character analysis. The various tasks performed on the image in the preprocessing stage are binarization, noise removal, and normalization etc as shown in Fig. 20.1.

In this section we illustrate the various methodologies involved in the preprocessing step. The document image is first binarized and sent for noise removal with suitable filters. We have divided the segmentation process in two phases. The first

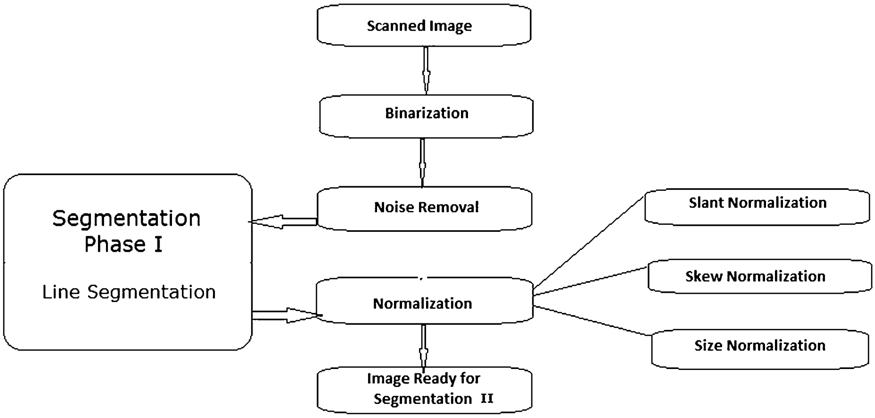


Fig. 20.1 Text document analysis: preprocessing steps

phase of segmentation is called line segmentation which is done as a part of preprocessing step. Line segmentation is done after binarization but before normalization, as the skew and slant normalization requires the segmented lines. Thus the preprocessing step takes scanned image of text document as input and gives output as segmented lines in normal form. Those segmented lines are sent to second phase of segmentation where words and and also characters are further separated for text digitization and hence recognition.

20.2.1 Binarization

The main goal of binarization is to convert a gray scale input image into a binary image, because many vision algorithms and operators only handle the binary image rather gray scale image. Selection of a global threshold value is a general technique to convert gray scale image into binary image. Such general methods binarize the entire image using a single threshold value, but this approach has many shortcomings as a global threshold value would have a side effect of excluding some text or inclusion of noise and vice-versa. Also finding a global threshold value is computationally complex. Basically binarization techniques are classified in two classes: global and adaptive. Various global binarization techniques in literature use intensity histogram calculation methods. One simple way is to automatically select the value at the valley of intensity histogram of the image, assuming that there are two peaks in the histogram, one corresponding to the foreground, the other to the background. Poor contrast and strong noise in the input image is also a challenge for selection of proper threshold value because due to the poor contrast, many images do not have such two peaks in the histogram. Various binarization approaches are used in handwritten text recognition system. In most of the systems researchers generally use the Otsu [3] method for binarization. In this section, a novel adaptive approach for text binarization is proposed which is particularly used for handwritten text document. Though it we can handle printed text as well. It is specifically very simple with constant complexity and uses only addition operation, hence it is very efficient for handwritten documents. For example, for banking and postal applications, like cheque detail recognition and postal address recognition etc. The technique uses the I_{max}-maximum and I_{min}-minimum intensity levels of a given handwritten document. We know that all minimum intensity pixels with some variations ($+I_O$) (intensity-offset) made by pen, are our foreground pixels and rest is background. Basically two facts are used for this novel approach: (a) for proper visualization or readability of the document a writer make a common sense in selecting a good pen. Which always distinctly differentiate between the background and foreground (handwritten content). If the background is dark then he will use a light pen and if the background is light then he will use a dark pen. So we assume that our text has a single and higher intensity text and rest part is assumed to be background. Taking our example further we can say that generally a writer writes or fill the cheque details in a cheque and fill the postal address in post card by a single pen; (b) when a user uses a single pen then the variation of intensity depends on the type of pen which is used by writer. The variation of intensity is also consider in our approach as average for both dark and light pen.

The intensity histogram computation based approaches [3–6], generally use two peaks for finding the threshold value, but many images do not have such two peaks in the histogram. The approach proposed have two version:

1. Global- for those handwritten documents where the text written in input image is having more or less the same intensity value or written by a single pen. Basically it calculates a global threshold. However a unique feature of this algorithm is that it is very simple and performs better than Otsu [3] method and Niblack's [4] method.
2. Adaptive-After global approach implementation we found that it is giving the best results for image with text written by single pen, but the result degrades if we use multiple pens. This happens due gradual intensity variations among the pens and drastic intensity variations across pens, which any global threshold measuring method would fail to address. Thus, we need an adaptive approach which could handle such huge intensity variations across a document.

To understand the binarization algorithms let us assume, that the input image has black foreground pixels and white background pixels (reverse it for the other way around). The global binarization method is illustrated in Algorithm 1

Algorithm 1 Global binarization.

```
Ensure: $I$- Binarized image, with background as 0.
Require: $I(i,j)$- Text document of gray level image, with $i \in M$ width and $j \in N$ height. {Find
  maximum $I_{max}$ and Minimum $I_{min}$ intensity value of pixels. Assuming that our background is
  with this global maximum intensity}.
  $I_{diff} = I_{max} - I_{min}$.
  if $(I_{diff} > I_{mean})$ then
    $I_o = 100$.
  else
    $I_o = 20$.
  end if
  $L_{th} = I_{min} + I_o$
```

Here, if difference between foreground and background intensity is large than average value than possible variation range of text content in also in large range and so offset set to be 100 and for small difference, variation range of text content also be small so offset set to be 20.

The adaptive approach is enumerated in Algorithm 2 [assuming, that the input image has black foreground pixels and white background pixels (reverse it for the other way around)].

Algorithm 2 Adaptive Binarization.

```
Ensure: $I$- Binarized image, with background as 0.
Require: $I(i,j)$- Text document of gray level image, with $i \in M$ width and $j \in N$ height. {Find
  global maximum $IG_{max}$ intensity value of pixels. Assuming that our background is with this
  global maximum intensity}.
  $IG_{max} = \max I(i,j)$.
  Split the input image into small blocks with size of $15 \times 15$.
  compute the minimum intensity $\forall blocks \mathrm{B}_{k,l}$.
  $IL_{min} = \min \in B_{k,l}$.
  $I_{diff} = IG_{max} - IL_{min}$.
  if ($I_{diff} > I_{mean}$) then
    $I_o = 100$.
  else
    $I_o = 20$.
  end if
  $L_{th} = IL_{min} + I_o$
```

Here, if difference between local foreground and global background intensity in block image is greater than average value than possible variation range of text content is also large and hence offset is set to be 100, and for small difference, the intensity variation range of text will also be small so offset is set to be 20. Experimental results show that the offset values of 20 for lower intensity variations and 100 for large variations could cater to general pen pressure and color variations, and hence gives very good results.

20.2.1.1 Experimental Results

We have used 100 test images (both handwritten, printed, single pen and multiple pen) for comparing our results. Fifty percent of the test images are downloaded from internet and rest are written by 20 writers on different quality of paper to address the problem of noise on different backgrounds. Our results are compared with standard algorithms like: Otsu [3], Niblack [4], Wu Manmatha [5], Srihari [6] for comparative analysis.

The adaptive approach when applied on the image having handwritten contents written through multiple pens with different coloured ink as shown in Fig. 20.2a; gives good results as shown in Fig. 20.2b.

The quantitative analysis regarding the computation cost are: our approach is comparatively fast than others because we are just finding the maximum and minimum intensity from our image and then applying addition operation for adding offset. All the others methods use histogram calculation followed by finding between class variations using multiple comparative operator for appropriate threshold selection.

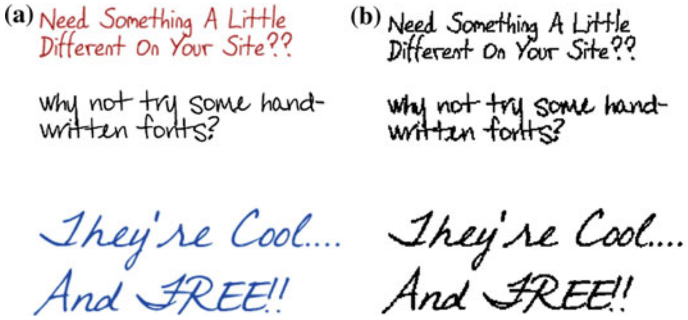
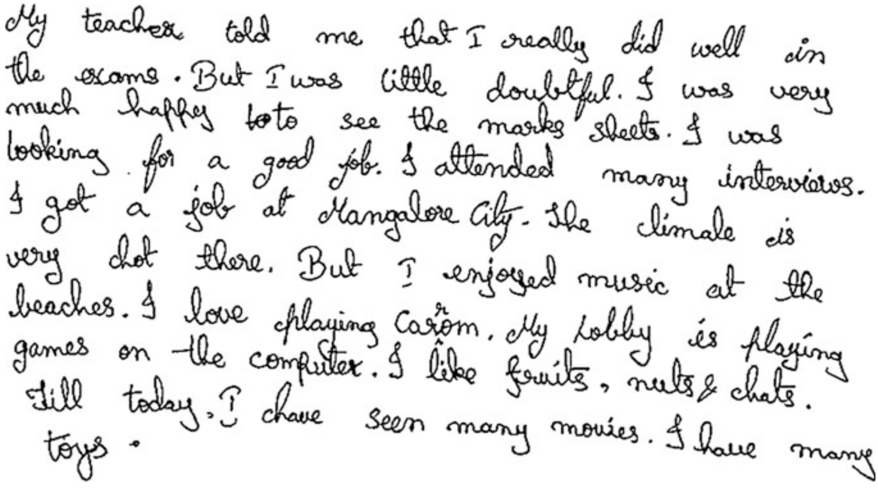

(a)
Need Something A Little
Different On Your Site??
why not try some hand-
written fonts?
They're Cool....
And FREE!!
(b)
Need Something A Little
Different On Your Site??
why not try some hand-
written fonts?
They're Cool....
And FREE!!

Fig. 20.2 **a** Example image of a handwritten text document with multiple pen intensities. **b** Binarization output using proposed adaptive approach

20.2.2 Line Segmentation

Text line segmentation of handwritten documents is much more difficult than that of printed documents. Unlike the printed documents which have approximately straight and parallel text lines, the lines in handwritten documents are often un-uniformly skewed and curved. Moreover, the spaces between handwritten text lines are often not obvious compared to the spaces between within-line characters, and some text lines may interfere with each other. Therefore many text line detection techniques, such as projection analysis [7] and K-nearest neighbor connected components (CCs) grouping [8], are not able to segment handwritten text lines successfully. Figure 20.3 shows an example of unconstrained handwritten document.



My teacher told me that I really did well in
the exams. But I was little doubtful. I was very
much happy ~~to~~to see the marks sheets. I was
looking for a good job. I attended many interviews.
I got a job at Mangalore City. The climate is
very hot there. But I enjoyed music at the
beaches. I love playing Carrom. My hobby is playing
games on the computer. I like fruits, nuts & chats.
Till today. I have seen many movies. I have many
toys.

Fig. 20.3 Example image of a general handwritten text paragraph

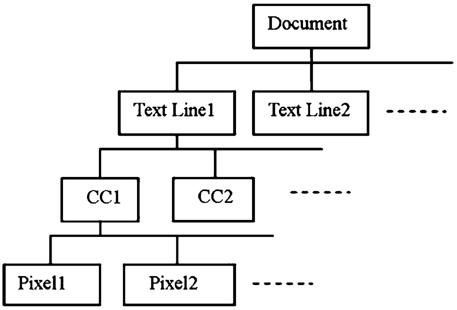


Fig. 20.4 Tree structure for a general text document

We could depict the relationship of text document components using a tree as shown in Fig. 20.4. Text document image segmentation can be roughly categorized into three classes: top-down, bottom-up, and hybrid. Top-down methods partition the document image recursively into text regions, text lines, and words/characters with the assumption of straight lines. Bottom-up methods group small units of image (pixels, *CC*s, characters, words, etc.) into text lines and then text regions. Bottom-up grouping can be viewed as a clustering process, which aggregates image components according to proximity and does not rely on the assumption of straight lines. Hybrid methods combine bottom-up grouping and top-down partitioning in different ways.

All the three approaches have their advantages and disadvantages. Top-down methods work well for typed text where the text lines are relatively horizontal, but it does not perform well on curved and overlapping text lines. The performance of bottom-up grouping relies on some heuristic rules or artificial parameters, such as the between-component distance metric for clustering. On the other hand, hybrid methods are complicated in computation, and the design of a robust combination scheme is non-trivial.

We are proposing an effective bottom-up grouping method for text line segmentation for unconstrained handwritten text documents. Our approach is based on minimal spanning tree (MST), grouping of *CC*s, and the connectivity strength function (CSF).

For line segmentation, we first extract the connected components from the image (binarized) using 8-connectivity for foreground pixels and 4-connectivity for background pixels. To construct a line from these connected components, calculate the centroid of every connected component as depicted in Fig. 20.5 with green colored pixels. Using these centroids of connected components as vertices (graph), calculate the cost matrix of the given graph, where the cost of the edge is the distance between two vertices. The minimum spanning tree is then calculated for the graph as shown in Fig. 20.5.

The *MST* as shown in Fig. 20.5, also have some mis-qualified words (example marked with yellow circle) linked with the line words which are not part of the same line. For removing such connections (edges) from the *MST*, we further use a connectivity strength function as explained below which is very useful in deciding

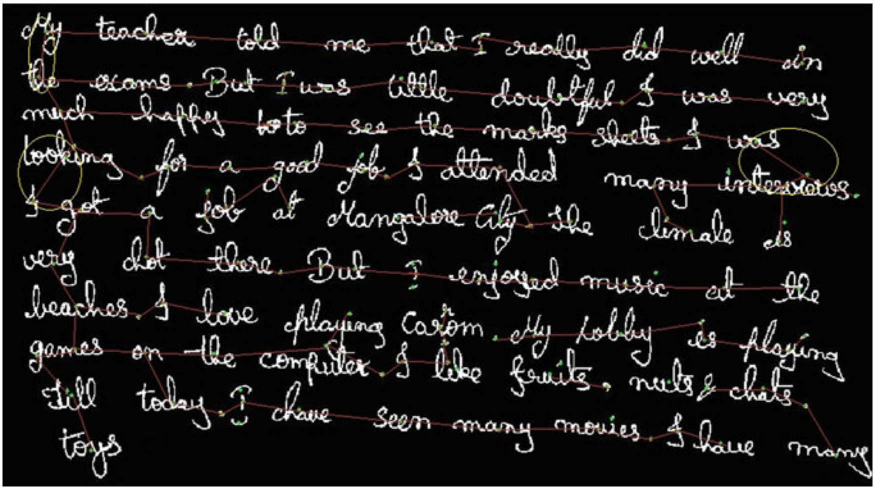

My teacher told me that I really did well in
the exams. But I was little doubtful. I was very
much happy to to see the marks sheets. I was
looking for a good job. I attended many interviews.
I got a job at Mangalore City. The climate is
very hot there. But I enjoyed music at the
beaches. I love playing Carom. My hobby is playing
games on the computer. I like fruits, nuts & chats
till today. I have seen many movies. I have many
toys

Fig. 20.5 *MST* generated for the text paragraph shown in Fig. 20.3, the green Pixels mark the centroids of every connected component, and the red lines depict the edges of the *MST* of the graph of the connected components of the same figure

the groups of the components which belongs to the same line. Thus the misaligned edges are removed from the *MST* and we generate the correct forest of the connected components.

The Connectivity Strength Function *CSF* is derived as, let there be two connected components C_1 and C_2 having centroids as (x_1, y_1) and (x_2, y_2), respectively. The minimum distance (d) between the two components is

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

and the vertical distance Y_d is

$$y_d = (y_2 - y_1)$$

then the *CSF* is defined as

$$CSF = \frac{\mid d - y_d \mid}{y_d}$$

For each pair of connected component in the *MST* compute the value of *CSF*. The decision for grouping the components depends on *CSF* as

$$CSF = \begin{cases} 0 & \text{belongs to different lines.} \\ \infty & \text{belongs to same line.} \end{cases}$$

where, $CSF = 0$, only when $d = y_d$, means the two components have the minimum connectivity strength as they are orthogonal and hence belongs to different lines, which are almost parallel. And the $CSF = \infty$, only when $y_d = 0$. This means

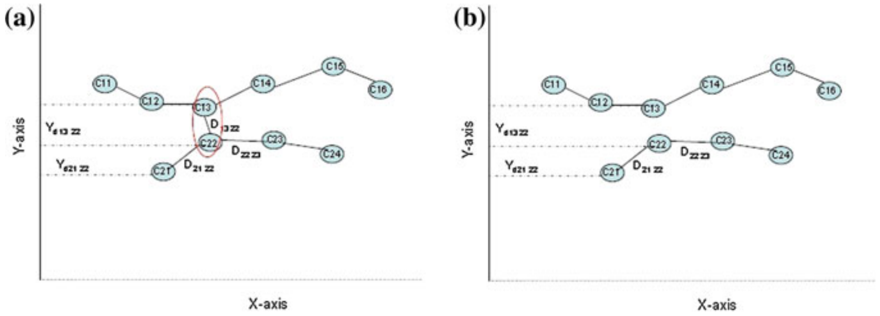


Fig. 20.6 Illustration of connectivity parameters: **a** *MST* with mis-qualified edge marked with red circle; **b** Forest generated after removing this mis-qualified edge using *CSF* on the *MST*

the connectivity between the two components is the strongest as they both belong to the same line. The angle between the two components is zero aligning them on the same line.

For example consider the *MST* as shown in Fig. 20.6. The labels c_{11}, c_{12}, c_{13}, c_{14}, c_{15}, c_{16} shows the connected components of same line and $c_{21}, c_{22}, c_{23}, c_{24}$ belongs to another line. In this example the connected component c_{22} has degree >3, and it is more closer to c_{13} than his own line components. We calculate the *CSF* of c_{22} with c_{13}, c_{21}, c_{23}, and found that c_{22} is weekly connected with c_{13} so we remove the edge (c_{22}, c_{13}) from the *MST* to show the correct *MST*.

Thus, after applying the *CSF* rules on Fig. 20.3 we remove the mis-aligned components from the text lines and generate the forest of given document image as shown in Fig. 20.7. Where our forest is defined as a group of trees. Where every tree is a text line.

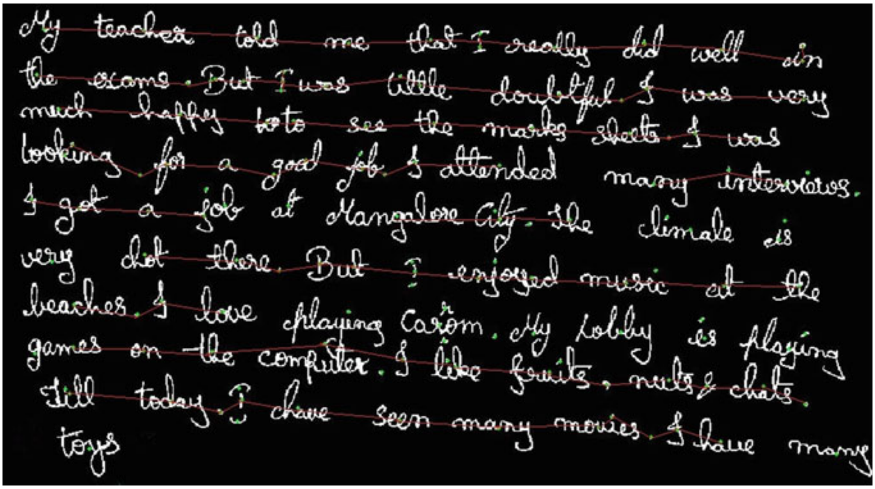

My teacher told me that I really did well in
the exams. But I was little doubtful. I was very
much happy to see the marks sheets. I was
looking for a good job. I attended many interviews.
I got a job at Mangalore City. The climate is
very hot there. But I enjoyed music at the
beaches. I love playing Carom. My hobby is playing
games on the computer. I like fruits, nuts & chats.
Till today I have seen many movies. I have many
toys

Fig. 20.7 Forest remains after removing the week edges from Fig. 20.3 using *CSF*

The complete process can be enumerated as shown in Algorithm 3.

Algorithm 3 Text Line Segmentation.

```
Ensure: $F$- Forest of text lines.
Require: $I$- Text document binarized image with background as 0.
  Compute connected components ($CC_i$)s using 8-connectivity.
  Compute centroids for all $CC_i$s.
  $\{(c_x, c_y) = (\frac{1}{M}\sum x_j, \frac{1}{M}\sum y_j)$, where $x_j, y_j \in CC_i$; $M$ is the number of pixels in $i^{th}$ $CC$.$\}$
  $\{$Compute cost matrix of $CC$s$\}$
  $d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$.
  Compute $MST$ of $G$ (graph) using cost matrix of $CC$s.
  $\{\forall V$(vertex) $\in G$with$V_{degree} \geq 2 \in MST$, apply $CSF$.$\}$
  if $(V_{degree} \geq 2)$ then
    $y_d = (y_2 - y_1)$.
    $CSF = \frac{|d - y_d|}{y_d}$.
    $\{\forall CC$ connected with this $V$.$\}$
  end if$\{$Remove the week edges where ever $CSF \approx 0$.$\}$
  Return($F$) $\{$Finally remains a forest having the trees for every single line.$\}$
```

20.2.2.1 Experimental Results

The experimental comparison of proposed line segmentation approach shows that proposed *CSF* improves text line segmentation accuracy significantly. The proposed method was also compared with other state-of-the-art methods in experiments on a large database of *IAM*, handwritten documents data set and its superiority was demonstrated.

The only side effect of using the *CSF* is over segmentation of line, but it is not a problem as the text lines are sequential so a line can be broken into two lines or more without effecting the syntactic or semantic meaning of the text. The advantage of *CSF* is it never mis-classifies a word with an incorrect text line. As a part of future work we can overcome the over segmentation through finding the sequential chain of trees of the forest.

20.2.3 Skew Normalization

The majority of both segmentation and character recognition algorithms are sensitive to the orientation of the word. Furthermore, the skewed words are very often found in handwritten text. Even in the case of correctly oriented pages, the handwritten words could present smaller or larger skews. Technically, skew is defined as the deviation of the base lines of the text from the horizontal direction. At the time of the recognition, handwritten text should be free from skew for better recognition rate and computational cost. It is due to the fact that, skew normalization is used to generate the uniform representation of particular character of given handwritten text image. In this paper , a simple and novel approach for Skew normalization for handwritten script is proposed. The concept of orthogonal projection with respect to x-axis is used in the skew detection. The orthogonal projection of the word image is maximum with un-skewed or normal text line or word image.

Algorithm 4 Text Line Image Skew Correction

```
Ensure: $I$-Skew corrected.
Require: $I$-Skewed image with intensity of background as 0.
  $P_l = P_{last} - P_{start} - \delta$ {Calculate the orthographic projection of text line on $x$-axis.}
  $I = I_{(R,+2)}$
  $P_{l_+} = I_{Orthographic}$ {Orthographic projection of rotated by $(+2^o)$ image}
  $I = I_{(R,-2)}$
  $P_{l_-} = I_{Orthographic}$ {Orthographic projection of rotated $(-2^o)$ image}
  Set $\theta = 1$ {Initialize $\theta$}
  if ($P_{l_+} > P_l$) then
      {Positive skew in image.}
      while ($P_{l_+} > P_l$) do
      $I = I_{(R,\theta)}$ Image rotated by angle $\theta$;
      $\theta + = 1$;
      $P_{l_+} = I_{Orthographic}$
    end while
  else if ($P_{l_-} > P_l$) then
      {Negative skew in image.}
      while ($P_{l_-} > P_l$) do
      $I = I_{(R,-\theta)}$ Image rotated by angle $-\theta$;
      $\theta + = 1$;
      $P_{l_-} = I_{Orthographic}$
    end while
  end if
  Return($I$)
```

The Fig. 20.8 shows that the $|OP|$ is the actual length (A_l) of line which is shown with the different skew angles about x-axis as $0, r', r''$ and 90 degrees. The relation between the A_l of segmented text line and projected length (P_l) with respect to skew angle θ is given as

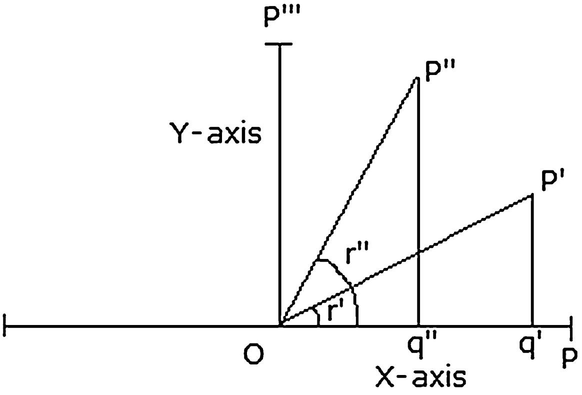


Fig. 20.8 Orthogonal projection of skewed text line

$$P_l = (A_l \times \cos\theta)$$

However in real scenario a segmented text line have height h and its projection δ is also considered in the projection calculation. The value of δ depends on the height of text line h,

$$\delta = h \times \cos(90 - \theta)$$

here, $\delta \in [0, h]$. (refer in Fig. 20.9)

The actual projection reduced by value of δ.

In Fig. 20.8 the $\delta = 0$ because $h = 0$

So we can say that,

$$OP = Oq' \times \cos r'$$

If skew angle θ is as $r' = 0$, then $\cos\theta = \cos r' = 1$, so

$$P_l = A_l$$

and hence, $OP = Oq'$

If Skew angle θ is, 90 *degree*, then $\cos 90 = 0$,

$$P_l = 0$$

The Fig. 20.8 show the orthogonal projection on the x-axis of each skewed line OP, OP', OP'', OP''', which are same as $|OP|$, is given by $OP, Oq', Oq'', 0$ respectively. Using this concept, we rotate the input segmented text line image till reach to maximum projected length.

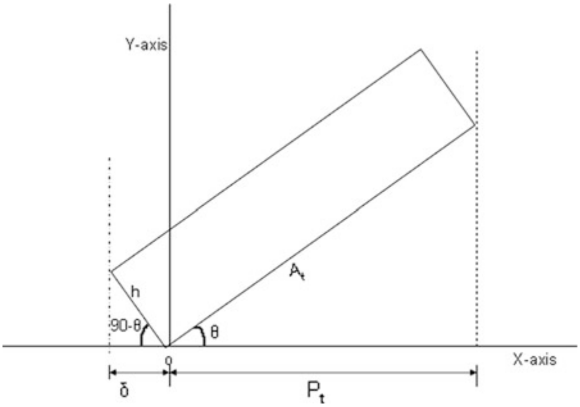


Fig. 20.9 Orthogonal projection of image of height h with effect of δ introduced in skew angle correction

Table 20.1 Results comparison

Compare methods	Skew angle (Degrees)	Accuracy (%)	Complexity
Bounding box method	0 – 25	78.40	$m \times n$
Linear regression	0 – 360	85.20	$m \times n$
Hough transformation	0 – 180	98.55	$m \times n$
Proposed method	0 – 360	98.30	Linear m

20.2.3.1 Results and Comparative Analysis

We have implemented the proposed method on a set of 500 handwritten text Images with resolution of 800 × 600 Pixels. The proposed method could determine the exact amount of skew in each image efficiently.

We have also implemented the Linear regression and bounding box methods, and Hough transform. The comparative analysis is summarized in Table 20.1.

20.3 Conclusions

It has been observed experimentally that all types of images could be successfully binarized by our proposed adaptive binarization approach and we get better results than others as discussed in Sect. 20.2.1. Our approach is also very fast as it uses only integer addition operation(s), compared to histogram computation by others. The proposed text line segmentation approach with the novel use of *CSF* has the advantage of ≈ 0 mis-classification rate of words to incorrect text lines. We have also proposed a new approach of skew normalization for handwritten text document with linear time complexity. The proposed approach determines the exact skew and is computationally very efficient compared to existing techniques, as it does simple comparison operations for skew detection. It could handle all skews (360°), and hence could be adapted to different language writing styles with equal accuracy. For example, we write left to right in most of the languages, but in Urdu language we write from right to left. All such cases could be handled by our approach, where rotation angle varies only from $[0 : 90]$.

References

1. Horst B (2003) Recognition of cursive Roman handwriting- past, peresent and future. Proceeding of the seventeen international conference on document analysis and recognition ICDAR, In, pp 448–459
2. Neeta N, Ankit A, Anshul T, Gourav J (2012) Neural network based cursive handwritting recognition, lecture notes in engineering and computer science. In: Proceedings of the World congress on engineering and computer science, WCECS 2012, 24–26 October, San Francisco, USA pp 692–698

3. Otsu N (January 1978) A threshold selection method from Gray levelhistogram. IEEE Trans Syst, Man, Cybern 19:62–66
4. Niblack W (1986) An introduction to digital image processing. Prentice-Hall, New Jersey
5. Wu V, Manmatha (Jan 1998) Document image clean-up and binarization. Proc SPIE Conf Document Recog 3:18–23
6. Liu Y, Srihari SN (May 1997) Document image binarization based on texture features. IEEE Trans PAMI 19(5):540–544
7. Zamora-Martinez F, Castro-Bleda MJ, Espaa-Boquera S, Gorbe-Moya J (2010) The 2010 international joint conference on unconstrained offline handwriting recognition sing connectionist character N-grams, neural networks (IJCNN) pp 1–7, 18–23July 2010.
8. Kumar M, Jindal MK, Sharma RK (2011) K-nearest neighbour based offline handwritten gurumukhi Character recognition. In: International IEEE conference on image, information processing(ICHP 2011) vol 1, pp 7–11.
9. Sarfraj M, Rasheed Z (2008) Skew estimation and correction of text using bounding box. Fifth IEEE conference on computer graphics, imaging and visualization, In, pp 259–264
10. Sarfraz M, Mahmoud SA, Rasheed Z (2007) On skew estimation and correction of text, computer graphics, imaging and visualization IEEE computer society USA pp 308–313.
11. Nagy G, Seth S, Viswanathan M (1992) A prototype document image analysis system for technical journals. Computer 25(7):10–22

Chapter 21
Tracing Malicious Injected Threads Using Alkanet Malware Analyzer

Yuto Otsuki, Eiji Takimoto, Takehiro Kashiyama, Shoichi Saito, Eric W. Cooper and Koichi Mouri

Abstract Recently, malware has become a major security threat to computers. Responding to threats from malware requires malware analysis and understanding malware behavior. However, malware analysts cannot spend the time required to analyze each instance of malware because unique variants of malware emerge by the thousands every day. Dynamic analysis is effective for understanding malware behavior within a short time. The method of analysis to execute the malware and observe its behavior using debugging and monitoring tools. We are developing Alkanet, a malware analyzer that uses a virtual machine monitor based on BitVisor. Alkanet can analyze malware even if the malware applies anti-debugging techniques to thwart analysis by dynamic analysis tools. In addition, analysis overhead is reduced. Alkanet executes malware on Windows XP, and traces

Y. Otsuki
Graduate School of Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
e-mail: yotuki@asl.cs.ritsumei.ac.jp

T. Kashiyama
Ritsumeikan Global Innovation Research Organization, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
e-mail: kashiyama@asl.cs.ritsumei.ac.jp

S. Saito
Graduate School of Engineering, Nagoya Institute of Technology,
Gokiso-cho, Showa-ku, Nagoya, Aichi 466-8555, Japan
e-mail: shoichi@nitech.ac.jp

Eiji. Takimoto · E. W. Cooper · K. Mouri (✉)
College of Information Science and Engineering, Ritsumeikan University,
1-1-1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
e-mail: mouri@cs.ritsumei.ac.jp

Eiji. Takimoto
e-mail: takimoto@asl.cs.ritsumei.ac.jp

E. W. Cooper
e-mail: cooper@is.ritsumei.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_21,

system calls invoked by threads. Therefore, the system can analyze malware that infects other running processes. Also, the system call logs are obtained in real time via a IEEE 1394 interface. Other programs can readily examine the log and process the analysis results to understand intentions of malware behavior. In this paper, we describe the design and implementation of Alkanet. We confirm that Alkanet analyzes malware behaviors, such as copying itself, deleting itself, and creating new processes. We also confirm that Alkanet accurately traces threads injected by malware into other processes.

Keywords Dynamic analysis · Malware's behavior · Malware analysis · System call tracing · Thread injection · Virtual machine monitor

21.1 Introduction

Recently, malware has become a major security threat on computers. According to a report released by Symantec Corporation, more than 403 million unique variants of malware were detected in 2011 [1]. This number was an increase of 41% over the previous year. Responding to this threat requires analysis and understanding of malware behavior. However, malware analysis cannot spend a lot of time on each malware because new variants emerge by the thousands every day.

The first step of malware analysis is a dynamic analysis to gain a understanding of the malware's general behavior. In this step, analysts execute the malware and observe its behavior using debuggers and monitoring tools. The next step targets only complicated or notable malware. In such cases, analysts do a detailed dynamic analysis or static analysis using disassemblers and debuggers to read the malware code. Malware analysts need to form a summary of malware in a short time because new malware is constantly emerging. We focus attention on the first dynamic analysis.

Dynamic analysis differs from static analysis in that it is not thwarted by techniques to interfere with program analysis such as packing and obfuscation. So dynamic analysis can provide summary reports of malware in a short time. However, recent malware has applied anti-debugging techniques [2, 3]. When malware detect that it has been analyzed by dynamic analysis tools, it may attempt to evade analysis or tamper with the analysis tools. Traditional dynamic analysis tools run using services provided by Windows to assist in debugging. However, these tools and services do not run stealthily, hidden from the program being debugged, because their purpose is to debug legitimate software. Therefore, malware can detect them easily. To analyze malware that applies anti-debugging techniques, these techniques need to be disabled one by one using a debugger or kernel-mode driver. Alternatively, analysts should use other analysis tools and not depend on general debugging assistant services. However, it is very difficult to evade all anti-debugging techniques. In addition, side effects to the environment in which the malware is executed are increased by falsifications needed to evade anti-debugging techniques, which cause increased overhead and inaccurate analysis.

Analysts need to use dynamic analysis to reduce side effects to environments for executing malware. Observing from outside the execution environment is an effective method of reducing the side effects. Some dynamic analysis system implementations have been based on virtual machine monitors (a.k.a. VMM) or emulators. VMMs and emulators run under a higher privilege level than the operating systems in virtual machines, which can observe all user-mode processes and the operating system in a virtual machine transparently. However, emulating the whole environment by software alone incurs a huge processing overhead. In addition, malware can detect general VMM easily. The reason is that general VMM emulates specific hardware and has communication interfaces with the guest operating system.

Consideration of the analysis targets reveals another problem. Recently, more and more malware has the capability of spreading outside the range of a single process. The purposes of such spreading behavior are to conceal activities and to make analysis difficult. Sophisticated malware injects malicious codes and threads into other processes. Existing dynamic analysis systems cannot trace malicious threads injected into legitimate processes because the systems distinguish a current executing task by process level.

It is also necessary to give consideration to granularity of analysis. Instruction level analysis can analyze particularly. However, it causes too much overhead and it is hard to interpret the intentions of malware behavior. An API trace that records Windows API functions called by malware gives abstract and useful information, and can be obtained in a short time. There is the possibility that the malware alters user-mode memory space to evade tracing. However, malware needs to invoke system calls to affect to the environment when it is running on user-mode. Therefore, system call tracing is an effective method for malware analysis.

We are developing the system a dynamic analysis system to meet the following requirements, based on the above-mentioned reasons.

1. The system can analyze malware that applies anti-debugging techniques.
2. The system can analyze malware that infects other running processes.
3. The system traces system calls invoked by malware.
4. The analysis method is one that mitigate analysis overhead.

To meet these requirements, we are developing Alkanet [4], a malware analyzer based on VMM. The remainder of this paper is organized as follows. Section 21.2 gives an overview of Alkanet. Section 21.3 details our system call tracing method. Section 21.4 presents malware analysis results and other considerations of using Alkanet. Section 21.5 summarizes related works. Finally, Sect. 21.6 gives the conclusions of this paper.

21.2 Overview of Alkanet

21.2.1 Outline

Alkanet is a dynamic analysis system for malware analysis using VMM. Malware analyzers implemented in VMM can analyze with a higher privilege level than malware. So many anti-debugging techniques would be ineffective for Alkanet. Alkanet can observe running malware without the malware interfering.

From two viewpoints, API level tracing is better suited for analyzing malware within a short time than instruction level analysis. One viewpoint is the ease of understanding intention of a behavior. The other is analysis overhead. In addition, malware running in user-mode needs to invoke system calls to affect the environment. Therefore, Alkanet traces invoked system calls and analyzes malware behaviors.

Achievement of system call tracing requires hooking every system call invoked by malware and getting the arguments and return value. In addition, the tracing requires analysis of the meaning of the arguments and return value to get detailed information of system calls invoked. However, a VMM cannot get abstract information on the operating system level. A VMM cannot call the API provided by operating system. The reason is that the VMM runs outside of the guest operating system. Many existing dynamic analysis systems using a VMM or emulator resolve this problem by receiving internal information from their agent processes or drivers running in Windows. However, this solution carries the risk that malware detects or tampers with the agent and the communication interfaces. A VMM or emulator can conceal or protect the agent and the communication interfaces from malware. These efforts cause an increase in analysis overhead. Our approach to the problem is that Alkanet itself refers to a memory region of Windows and obtains detailed information of Windows. VMM can access all memory regions of the virtual machine containing the operating system because VMM runs at a higher privilege level than the guest operating system.

Malware behaviors can be analyzed from the system call logs. However, the system call tracer in Alkanet records a high volume of logs in cases where the analysis target is large-scale or sophisticated malware. For these reasons, our system extracts a summary of malware behavior from further analysis of the system call logs.

21.2.2 Overview

Figure 21.1 presents the overview of Alkanet. Alkanet is implemented based on BitVisor [5]. BitVisor runs directly on the hardware and does not require a host operating system, and instead runs on processors with Intel Virtualization Technology (a.k.a. Intel VT). Intel VT assists virtualization by VMM. Therefore,

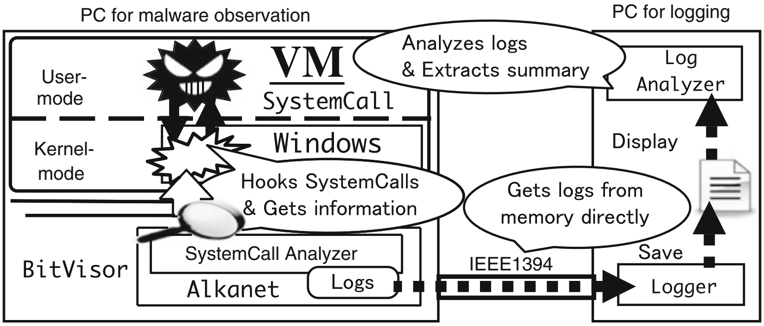


Fig. 21.1 Overview of Alkanet

BitVisor runs faster than emulators and VMMs implemented in software only. BitVisor can run Windows without requirement of modifications. In addition, BitVisor adopts the parapass-through architecture and does not emulate specific hardware. BitVisor provides the physical hardware for a guest operating system. Therefore, malware cannot detect BitVisor by characteristics of hardware, unlike emulators and VMMs that emulate specific hardware. Furthermore, Alkanet adopts Windows XP 32bit edition as its guest operating system. Alkanet executes malware in this environment. Alkanet hooks invoked system calls in this environment and records the number of system calls, their arguments, return values, and so on.

Another machine obtains the system call logs via IEEE 1394 interface. IEEE 1394 has direct read and write access to the physical memory of connected devices. This direct access allows the tracing logs to be obtained without the malware detecting or interfering. Other programs can readily examine the log and process the analysis results to understand intentions of malware behavior.

21.2.3 Observed System Calls and Getting Information

Table 21.1 gives an example of behaviors and system calls that Alkanet observes. Typical malware functions consist of these system calls.

Table 21.1 System calls observed by Alkanet

Behaviors	Examples of system calls
File	NtCreateFile, NtReadFile, NtWriteFile
Registry	NtCreateKey, NtQueryValueKey, NtSetValueKey
Network	NtDeviceIoControlFile, NtReadFile, NtWriteFile
Process	NtCreateProcessEx, NtTerminateProcess
Driver	NtLoadDriver, NtUnloadDriver
Code injection	NtCreateThread, NtWriteVirtualMemory

We have to distinguish a system call invoker by means of thread level because the execution unit in Windows is a thread. In addition, there are malicious threads in legitimate processes by code injection. Therefore, to distinguish a system call invoker, Alkanet obtains the Cid (pair that includes process id and thread id) and image name. In addition, it obtains system call arguments and return value in order to analyze malware behaviors. However, raw arguments and return value lack enough information for analysis because they consist of pointers and Windows-specific data structures. Therefore, the necessary information for analysis is obtained by interpreting these data structures. In the process described above, Alkanet obtains the following information.

- System call invoker's Cid and image name
- System call number
- System call arguments and return value
- Complementary information for Windows specific data structures.

21.3 System Call Tracing

21.3.1 System Call Hooking

A system call in Windows XP 32bit edition usually uses the sysenter and sysexit instructions. Sysenter enters from user-mode to kernel-mode. Sysexit returns from kernel-mode to user-mode. To get inputs given to system calls and their results, Alkanet hooks both sysenter and sysexit.

Figure 21.2 presents the flow of system call hooking by Alkanet. The following steps detail the flow.

1. Malware invokes a system call.
2. At the entry point of kernel-mode, transition from Windows to Alkanet occurs by a breakpoint.
3. Alkanet gets the necessary information.
4. Alkanet returns control to Windows.
5. Windows executes kernel functions.
6. At the exit point of kernel-mode, transition from Windows to Alkanet occurs by breakpoint again.
7. Alkanet gets the necessary information containing system call results.
8. Alkanet returns control to Windows.
9. Windows returns control to malware.

Alkanet uses hardware breakpoints to hook system calls. Alkanet sets breakpoints on entry point of KiFastCallEntry and exit point of KiSystemCallExit2. KiFastCallEntry is sysenter destination. KiSystemCallExit2 contains sysexit. These symbols are public by Microsoft [6]. We can get address of these symbols.

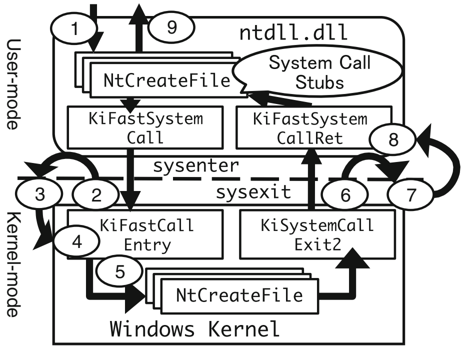


Fig. 21.2 Flow of system call hooking

Some anti-debugging techniques applied by malware detect hardware breakpoints. Countering such techniques requires concealment of the hardware breakpoints used by Alkanet. A VMM running on Intel VT can catch events in which a guest operating system has modified debug registers. So, Alkanet can conceal hardware breakpoints with a low overhead.

A system call consists a pair of sysenter and sysexit calls. We need to associate these logs. However, this does not mean that these logs are always consecutive. Therefore, Alkanet hooks each point individually. In log analysis phase, our system associates these logs using information of the system call number and invoker.

21.3.2 Identifying Invoked System Call

When a system call is invoked, Windows sets the system call number in the EAX register. Alkanet can get the system call number from the EAX register when hooking sysenter. On the other hand, the value of the EAX register has already been changed when hooking sysexit. System calls are invoked via stubs implemented in ntdll.dll. These stubs each have the same name symbol as their library function. For example, NtCreateFile system call is invoked via the NtCreateFile function of ntdll.dll. Therefore, Alkanet can identify invoked system calls by the return address pushed on the stack.

21.3.3 Identifying Invoker Process and Thread

When Alkanet hooks a system call invoked on Windows, Alkanet gets information of the process that invoked the system call. Windows has data structures of each processor state, called Processor Control Region (a.k.a. PCR) and Processor Control Block (a.k.a. PRCB). These data structures map on to the FS segment for each processor. The Windows kernel uses the structures, which have the address of the thread object running currently. Therefore, Alkanet can get information of the invoking process from the thread object referred by these data structures.

A thread object on Windows has a process id and thread id, a pair called Cid. Each thread object has also pointer to the process object the thread belongs to. The process object has the image name of the process. Therefore, Alkanet can get the Cid and image name of the process from the thread object and the process object.

21.3.4 Getting Return Values and Arguments

Windows APIs store return values to the EAX register and store arguments to the stack. Alkanet gets the return value from the EAX register and arguments from the stack. System calls in Windows save the value of the ESP register at the time to the EDX register to give the top of the stack in user-mode to the Windows kernel. The sysexit instruction loads the value of the ECX register into the ESP register. Alkanet gets the top of the stack in user-mode from the EDX register when hooking sysenter and from the ECX register when hooking sysexit.

Raw arguments and return values lack information to analyze because they consist of pointers and Windows specific data structures. Alkanet supplies the required information to analysis by referring to the memory region of Windows and interpreting these data structures. For example, NtCreateFile is a system call to create or open a file. Figure 21.3 shows the NtCreateFile declaration. We need to understand which files malware has attempted to open or create so it is necessary to get the file path that is passed to NtCreateFile. The third argument of NtCreateFile is a pointer to an OBJECT_ATTRIBUTES structure. The structure is used to set attributes of a Windows internal object. It contains the UNICODE_STRING type field, called ObjectName. In the case of NtCreateFile, the field is a unicode string for the file path. The third argument is read only, indicated by the argument annotation `__in` [7]. Therefore, Alkanet gets the arguments in both hooks.

Obtaining more detailed information about the the file requires referring to the corresponding file object. Windows manages resources (files, registries, processes,

Fig. 21.3 Declaration of NtCreateFile [28]

```
NTSTATUS NtCreateFile(
  __out     PHANDLE FileHandle,
  __in      ACCESS_MASK DesiredAccess,
  __in      POBJECT_ATTRIBUTES ObjectAttributes,
  __out     PIO_STATUS_BLOCK IoStatusBlock,
  __in_opt  PLARGE_INTEGER AllocationSize,
  __in      ULONG FileAttributes,
  __in      ULONG ShareAccess,
  __in      ULONG CreateDisposition,
  __in      ULONG CreateOptions,
  __in      PVOID EaBuffer,
  __in      ULONG EaLength
);
```

etc.) as objects. Each user-mode process has a handle table to manage objects opened by the process. A user-mode process interacts with resources using the handle corresponding to the object. In the case of NtCreateFile, the first argument is a pointer to a variable to receive a handle corresponding to the created or opened file object. The file object contains information about the file. For example, the object has a UNICODE_STRING type field that has the file path, called FileName. Alkanet refers to the memory region of Windows and gets detailed information of objects when the need arises. However, in the NtCreateFile case, the first argument is annotated with `__out`. `__out` arguments are written into by Windows kernel functions. Therefore, Alkanet gets these only when sysexit hooking.

21.4 Evaluation

21.4.1 Analysis Target Samples and Evaluation Methods

To confirm that Alkanet is effective for real malware analysis, we analyzed real malware samples using Alkanet. The samples are actual instances of malware recorded in CCC DATAset 2011 [8]. Here, we call the samples SdBot.exe, Palevo.exe and Polipos.exe based on the names assigned by some anti-virus software.

We executed these malware samples, traced invoked system calls, and analyzed the logs. We checked the validity of our analysis results by comparison with reports on the malware by anti-virus vendors. In this regard, however, there are large number of variants for the each of the malware and the variants differ from each other in the details. The two malware samples are not necessarily the same instance even if they have been detected as the same name by anti-virus software. It is hard to fully match the malware behaviors actually observed to reports by anti-virus vendors. Therefore, we evaluated whether our system could observe malware behavior characteristics in common with reports from several anti-virus vendors. In addition, we also confirmed whether there are variants of the malware that exhibit minor behaviors observed by Alkanet.

In this evaluation, we did not connect our system to networks. The reason is that Alkanet does not filter any network activity in its current implementation. We must prevent the malware being analyzed from performing actual attacks against real computers and servers.

21.4.2 SdBot

Figures 21.4, 21.5, 21.6 and 21.7 present a portion of the trace logs for SdBot.exe. The meaning of each item in the log entry is as follows.

No. & Time   Log number and CPU time for identifying log entries

Fig. 21.4 SdBot.exe: copying file of itself to system32 folder

```
No. : 1356     Time: 1697432035
Cid : 158.544  Name: SdBot.exe
Type: sysenter SNo.: b7 (NtReadFile)
Note: \...\My Documents\SdBot.exe

No. : 1357     Time: 1697432105
Cid : 158.544  Name: SdBot.exe
Type: sysexit  SNo.: b7 (NtReadFile)
Ret : 0 (STATUS_SUCCESS)
Note: \...\My Documents\SdBot.exe

No. : 4796     Time: 1697997485
Cid : 158.544  Name: SdBot.exe
Type: sysenter SNo.: 25 (NtCreateFile)
Note: \??\C:\WINDOWS\system32\ssms.exe

No. : 4800     Time: 1697998049
Cid : 158.544  Name: SdBot.exe
Type: sysexit  SNo.: 25 (NtCreateFile)
Ret : 0 (STATUS_SUCCESS)
Note: \WINDOWS\system32\ssms.exe

No. : 4806     Time: 1697999212
Cid : 158.544  Name: SdBot.exe
Type: sysenter SNo.: 112 (NtWriteFile)
Note: \WINDOWS\system32\ssms.exe

No. : 4807     Time: 1697999802
Cid : 158.544  Name: SdBot.exe
Type: sysexit  SNo.: 112 (NtWriteFile)
Ret : 0 (STATUS_SUCCESS)
Note: \WINDOWS\system32\ssms.exe
```

Fig. 21.5 SdBot.exe: deleting file of itself

```
No. : 8652      Time: 1700450001
Cid : 65c.2c0   Name: ssms.exe
Type: sysenter  SNo.: e0 (NtSetInformationFile)
Note: DELETE: \...\My Documents\SdBot.exe

No. : 8653      Time: 1700450168
Cid : 65c.2c0   Name: ssms.exe
Type: sysexit   SNo.: e0 (NtSetInformationFile)
Ret : 0 (STATUS_SUCCESS)
Note: DELETE: \...\My Documents\SdBot.exe
```

Fig. 21.6 SdBot.exe: setting to start automatically when system is rebooted

```
No. : 8656      Time: 1700451656
Cid : 65c.2c0   Name: ssms.exe
Type: sysenter  SNo.: f7 (NtSetValueKey)
Note: \REGISTRY\...\WINDOWS\CURRENTVERSION\RUN

No. : 8657      Time: 1700452279
Cid : 65c.2c0   Name: ssms.exe
Type: sysexit   SNo.: f7 (NtSetValueKey)
Ret : 0 (STATUS_SUCCESS)
Note: \REGISTRY\...\WINDOWS\CURRENTVERSION\RUN
```

Fig. 21.7 SdBot.exe: scanning device files used by dynamic analysis tools

```
No. : 9391      Time: 1700652627
Cid : 65c.678   Name: ssms.exe
Type: sysexit   SNo.: 25 (NtCreateFile)
Ret : c0000034 (STATUS_OBJECT_NAME_NOT_FOUND)
Note: \??\SICE

No. : 9415      Time: 1700746367
Cid : 65c.678   Name: ssms.exe
Type: sysexit   SNo.: 25 (NtCreateFile)
Ret : c0000034 (STATUS_OBJECT_NAME_NOT_FOUND)
Note: \??\REGMON

No. : 9417      Time: 1700762003
Cid : 65c.678   Name: ssms.exe
Type: sysexit   SNo.: 25 (NtCreateFile)
Ret : c0000034 (STATUS_OBJECT_NAME_NOT_FOUND)
Note: \??\FILEMON
```

Cid & Name	Information of system call invoker
Type & SNo.	Sysenter type or sysexit type and information of invoked
Ret	Return value (only sysexit entries)
Note	Additional information, for example about arguments

SdBot.exe ran as a process with id 158. The SdBot.exe process read its own file (No. 1356, 1357), created a file named ssms.exe in `C:\WINDOWS\System32` folder (No. 4796, 4800) and wrote the file (No. 4806, 4807). This behavior is presented by Fig. 21.4. After the behavior, the SdBot.exe process executed ssms.exe. The ssms.exe process deleted the original file of SdBot.exe (Fig. 21.5).

We confirmed that the ssms.exe process sets specific registry keys. Figure 21.6 shows that the ssms.exe process sets the Run key. Applications registered with the Run key start automatically when user logs onto the system. We confirmed that ssms.exe was registered with that key.

The ssms.exe process scanned device files regularly used by analysis tools. This behavior is one of general anti-debugging techniques. Figure 21.7 presents some sysexit entries of the recorded behaviors. Some dynamic analysis tools create specific device files. Therefore, malware can detect the tools by confirming the existence of specific device files. This kind of anti-debugging technique is ineffective for Alkanet because Alkanet does not create specific device files. In addition, Alkanet confirmed other malware behaviors such as executing cmd.exe and regedit.exe, creating and deleting temporary files, and modifying settings of services and networks.

Malware detected as SdBot by anti-virus software are backdoor Trojans [9, 10]. They await commands by attackers in internet relay chat (a.k.a. IRC). Additionally, there are many variants of SdBot. The infection method of SdBot is that the malware copies a file of itself to the Windows system folder. Each file copied then has a similar name to a regular Windows execution file. SdBot sets the copied file in Run key so that it will start automatically when the system has restarted.

In this analysis, we conclude that our system successfully analyzed the infection process of SdBot. It is because, as previously mentioned, our system could observe the behaviors such as Sdbot.exe dropping ssms.exe in the system32 folder and setting it in the Run key. It coincides with characteristics of SdBot that name of the ssms.exe is similar name of a Windows regular process, called smss.exe (Session Manager Subsystem). In addition, we confirmed that SdBot.exe modified network settings. However, we could not analyze the behaviors using IRC in detail because our analysis system was not connected to the network.

21.4.3 Palevo

Figures 21.8, 21.9 and 21.10 present part of the trace logs for Palevo.exe. Palevo.exe ran as a process with id 534. Figure 21.8 presents a behavior in which the malware restarted itself promptly. This behavior is an anti-debugging technique to

Fig. 21.8 Palevo.exe: restarting itself

```
No. : 3945      Time: 161921464
Cid : 534.538   Name: Palevo.exe
Type: sysenter  SNo.: 30 (NtCreateProcessEx)
Note: \...\My Documents\Palevo.exe

No. : 3946      Time: 161922097
Cid : 534.538   Name: Palevo.exe
Type: sysexit   SNo.: 30 (NtCreateProcessEx)
Ret : 0 (STATUS_SUCCESS)
Note: PID: 53c, ProcessName: Palevo.exe

No. : 3971      Time: 161926922
Cid : 534.538   Name: Palevo.exe
Type: sysenter  SNo.: 101 (NtTerminateProcess)
Note: PID: 534, ProcessName: Palevo.exe

No. : 3972      Time: 161926935
Cid : 534.538   Name: Palevo.exe
Type: sysexit   SNo.: 101 (NtTerminateProcess)
Ret : 0 (STATUS_SUCCESS)
Note: PID: 534, ProcessName: Palevo.exe
```

Fig. 21.9 Palevo.exe: copy file of itself to recycle bin of non-existent user

```
No. : 3988      Time: 161934677
Cid : 53c.540   Name: Palevo.exe
Type: sysenter  SNo.: b7 (NtReadFile)
Note: \...\My Documents\Palevo.exe

No. : 3989      Time: 161934724
Cid : 53c.540   Name: Palevo.exe
Type: sysexit   SNo.: b7 (NtReadFile)
Ret : 0 (STATUS_SUCCESS)
Note: \...\My Documents\Palevo.exe

No. : 4128      Time: 161959968
Cid : 53c.540   Name: Palevo.exe
Type: sysenter  SNo.: 112 (NtWriteFile)
Note: \RECYCLER\S-1-5-21-0243...\psyjo3.exe

No. : 4129      Time: 161960277
Cid : 53c.540   Name: Palevo.exe
Type: sysexit   SNo.: 112 (NtWriteFile)
Ret : 0 (STATUS_SUCCESS)
Note: \RECYCLER\S-1-5-21-0243...\psyjo3.exe
```

Fig. 21.10 Palevo.exe: thread injection to explorer.exe

```
No. : 4158      Time: 161976728
Cid : 53c.540   Name: Palevo.exe
Type: sysenter  SNo.: 115 (NtWriteVirtualMemory)
Note: PID: f8, ProcessName: explorer.exe

No. : 4159      Time: 161976811
Cid : 53c.540   Name: Palevo.exe
Type: sysexit   SNo.: 115 (NtWriteVirtualMemory)
Ret : 0 (STATUS_SUCCESS)
Note: PID: f8, ProcessName: explorer.exe

No. : 4164      Time: 161977822
Cid : 53c.540   Name: Palevo.exe
Type: sysenter  SNo.: 35 (NtCreateThread)
Note: PID: f8, ProcessName: explorer.exe

No. : 4165      Time: 161977934
Cid : 53c.540   Name: Palevo.exe
Type: sysexit   SNo.: 35 (NtCreateThread)
Ret : 0 (STATUS_SUCCESS)
Note: Cid: f8.544, ProcessName: explorer.exe
```

evade debugger attachment. General debugger cannot be attached to more than one process. Therefore, the malware tried restarting itself and leaving the debugger behind. The technique is ineffective for Alkanet because Alkanet hooks all invoked system calls and traces processes spawned by malware.

Figure 21.9 presents a part of the behaviors by new Palevo.exe process with id 53c. The new process read its file (No. 3988, 3989) and wrote psyjo3.exe in recycle bin of non-existent user (No. 4128, 4129). This behavior is the malware making copies file of its own files. The psyjo3.exe was registered to restart when system would be restarted.

In addition, Fig. 21.10 presents that Palevo.exe infected to other process. Palevo.exe wrote to memory space of explorer.exe (No. 4158, 4159) and created a thread in the process (No. 4164, 4165). The purpose of the behavior is to conceal its malicious thread in explorer.exe and its main threats. Palevo.exe process exited shortly after the behavior. We confirmed suspicious behaviors by the malicious thread in explorer.exe. For example, the malicious thread accessed network devices and set some registry values managing applications associated with file extensions.

Malware detected as Palevo by anti-virus software are bots forming the Mariposa botnet [11]. The behavior characteristics of Palevo include dropping copies to user folders or system folders. Some of variants drop copies to recycle bins [12, 13]. The behavior characteristics also include connecting to remote servers by its malicious thread concealed in explorer.exe [12, 14]. In addition, there are also variants containing the behaviors setting the registry values managing applications associated with file extensions [14]. These behaviors were observed in this evaluation. Infection methods of Palevo are through peer-to-peer networks or removable medias. However, we could not confirm the behaviors in detail because our system was not connected to networks and removable medias in this evaluation.

21.4.4 Polipos

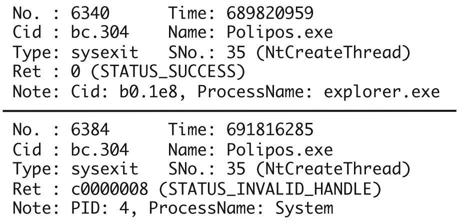
Figures 21.11, 21.12 present a part of trace logs for Polipos.exe. Figure 21.13, 21.14 present a part of results of log analysis. Polipos.exe process ran as a process with id 54c. Figure 21.11 shows that the Polipos.exe process has spawned new Polipos.exe process using NtCreateProcessEx. The new Polipos.exe process ran as a process with id bc.

Two sysexit entries in Fig. 21.12 presents that Polipos.exe process tried to inject a thread to other processes using NtCreateThread. The upper sysexit entry in Fig. 21.12 shows that Polipos.exe process with id bc injected a thread to an

Fig. 21.11 Polipos: swanning process

```
No. : 5786      Time: 677232506
Cid : 54c.6cc   Name: Polipos.exe
Type: sysenter  SNo.: 30 (NtCreateProcessEx)
Note: \...\My Documents\Polipos.exe

No. : 5787      Time: 677233114
Cid : 54c.6cc   Name: Polipos.exe
Type: sysexit   SNo.: 30 (NtCreateProcessEx)
Ret : 0 (STATUS_SUCCESS)
Note: PID: bc, ProcessName: Polipos.exe
```



```
No. : 6340      Time: 689820959
Cid : bc.304    Name: Polipos.exe
Type: sysexit   SNo.: 35 (NtCreateThread)
Ret : 0 (STATUS_SUCCESS)
Note: Cid: b0.1e8, ProcessName: explorer.exe
```

```
No. : 6384      Time: 691816285
Cid : bc.304    Name: Polipos.exe
Type: sysexit   SNo.: 35 (NtCreateThread)
Ret : c0000008 (STATUS_INVALID_HANDLE)
Note: PID: 4, ProcessName: System
```

Fig. 21.12 Polipos: thread injections

explorer.exe process. The malicious thread in explorer.exe was injected then ran with thread id 1e8. The thread copied the Polipos.exe file, and tried to connect to networks. The lower entry in Fig. 21.12 shows that Polipos.exe process tried to inject a thread into the System process using NtCreateThread. However, this NtCreateThread was failed because the return value was STATUS_INVALID_HANDLE.

Polipos.exe created threads in some other processes such as svchost.exe, service.exe, winlogon.exe, alg.exe, rundll32.exe, sqlservr.exe and lsass.exe. Figure 21.13 shows the thread tree. It was generated by analyzing system call logs and also by tracing threads that derived from other injected threads. In the figure, the Polipos's thread (Cid: 54c.18c) created new thread (Cid: 480.2c4) into svchost.exe (process id: 480). This thread (Cid: 480.2c4) created a new thread (Cid: 480.22c). This thread (Cid: 480.22c) created many new threads whose Cids are 480.720, 480.24c, and 480.7e0.

According to the thread (Cid: 480.720), code injection to rundl32.exe was found. It has created a new thread (Cid: 220.7f8). Alkanet could trace other derived threads from the thread (Cid: 220.7f8). Similarly, we found that the malicious thread on svchost.exe (Cid: 480.22c) also injected threads into other processes, such as explorer.exe and alg.exe. In this way, Alkanet can trace threads injected to other processes and derived from it also.

We focused attention on a Cid 480.41c thread. Figure 21.14 shows that the thread tried to open files such as drwebase.vdb, avg.avi, vs.vsn, anti-vir.dat. We recognized from return values that there were not these files. These files are used by some anti-virus software. Therefore, the behavior is a part of behavior deleting these files to prevent the malware itself being detected by anti-virus software.

Malware detected as Polipos or Polip by anti-virus software infects running processes [15, 16]. The malware infects most of the running processes and conceal its own existence. In addition, the malware also have a behavior of deleting specific files of anti-virus software. We confirmed these behaviors from analysis results by Alkanet as previously mentioned. Infection methods of Polipos are way to use Gnutella peer-to-peer networks. However, we could not confirm the behaviors in detail. It is because our system was not connected to networks in this evaluation.

```
No. [5212, 5213]: Polipos.exe (54c.18c) -> svchost.exe (480.2c4) (Code Injection)
 No. [5288, 5289]: svchost.exe (480.2c4) -> svchost.exe (480.22c)
  ...
  No. [11340, 11341]: svchost.exe (480.22c) -> svchost.exe (480.720)
   No. [14368, 14369]: svchost.exe (480.720) -> rundll32.exe (220.7f8) (Code Injection)
    No. [14546, 14547]: rundll32.exe (220.7f8) -> rundll32.exe (220.488)
     ...
  No. [11844, 11845]: svchost.exe (480.22c) -> svchost.exe (480.24c)
   No. [15080, 15081]: svchost.exe (480.24c) -> alg.exe (34c.1c8) (Code Injection)
    No. [15240, 15241]: alg.exe (34c.1c8) -> alg.exe (34c.5ac)
    ...
  No. [13214, 13215]: svchost.exe (480.22c) -> svchost.exe (480.7e0)
   No. [16586, 16587]: svchost.exe (480.7e0) -> explorer.exe (538.510) (Code Injection)
    No. [16744, 16745]: explorer.exe (538.510) -> explorer.exe (538.6ac)
    ...
```

Fig. 21.13 Polipos: thread tree generated by tracing a injected thread and derived threads from it

```
svchost.exe (480.41c)
 ...
 [NOT FOUND] No. [15246, 15247]: NtOpenFile \??\c:\program files\...\drwebase.vdb
 [NOT FOUND] No. [15248, 15249]: NtOpenFile \??\c:\program files\...\avg.avi
 [NOT FOUND] No. [15250, 15251]: NtOpenFile \??\c:\program files\...\vs.vsn
 [NOT FOUND] No. [15252, 15253]: NtOpenFile \??\c:\program files\...\anti-vir.dat
 ...
```

Fig. 21.14 Polipos: scanning files used by anti-virus software

21.5 Related Works

Olly Advanced [17] and Phant0m [18] are plug-ins of OllyDbg debugger [19] to disable many anti-debugging techniques and,VAMPiRE [20] makes it possible to use breakpoints without debugging assistant services. However, it is very difficult to evade all anti-debugging techniques as long as the tools run in the same environment as malware. It is also hard for a general debugger to analyze malware that infects other processes because the debugger is only attached to one process.

TTAnalyze [21] and its successor project Anubis [22, 23] are automated dynamic analysis systems based on QEMU [24]. However, malware can detect QEMU easily because QEMU emulates specific hardware. Malware cannot detect Alkanet because Alkanet does not emulate specific hardware. In addition, TTAnalyze cannot trace malicious threads in legitimate processes. Alkanet can trace malicious threads because Alkanet distinguishes malware by means of thread level. Furthermore, the QEMU emulation is inefficient because of the overhead required to emulate the whole environment in software only. Alkanet runs faster using hardware acceleration.

Virt-ICE [25] is a stealth debugger that extends and fixes QEMU. It is an effective interactive debugger for malware analysis. Ether [26] is a framework for malware analysis based on Xen [27]. Analysts can implement arbitrary malware analysis components using the framework. In other words, analysts need to implement malware analysis components using the framework to analyze malware actually.

21.6 Conclusion

In this paper, we describe “Alkanet”, a malware analyzer based on BitVisor. Our system can analyze malware within a short time even when the malware applies anti-debugging techniques to evade analysis by dynamic analysis tools. In addition, the analysis overhead is reduced. Alkanet executes malware on Windows XP, and traces system calls invoked by threads. The system call logs are obtained in real time via IEEE 1394. Therefore, the system can successfully analyze malware that infects other running processes. In addition, other programs can readily examine the log and process the analysis results to understand the intentions of malware behavior. We confirmed that Alkanet correctly analyzes malware behaviors (e.g. copying itself, deleting itself, creating new processes). We also confirmed that Alkanet precisely traces threads injected to other processes by malware.

In future work, we will make it possible to analyze a multitude of other methods to infect running other processes besides thread injection. We will also implement functions to make detailed observations of network behaviors. In addition, we will evaluate the possibility that the trace logs of Alkanet are practical for existing anomaly detection or malware clustering methods using system call logs.

References

1. Wood P et al. (2012) Internet security threat report vol 17 Symantec corporation, Tech rep
2. Falliere N (2007) Windows anti-debug reference. (2012) http://www.symantec.com/connect/articles/windows-anti-debug-reference Last accessed July 2012
3. Yason MV (2007) The art of unpacking. Black Hat USA.
4. Otsuki Y et al. (2012) Alkanet: a dynamic malware analyzer based on virtual machine monitor.In: Lecture notes in engineering and computer science: Proceedings of the World congress on engineering and computer science, WCECS 2012, vol 1 San Francisco, USA pp 36–44
5. Shinagawa T et al. (2009) BitVisor: a thin hypervisor for enforcing i/o device security.In: Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on virtual execution environments, ACM, Washington, DC, USA pp 121–130
6. Microsoft: standalone and remote debugging tools, Symbols, and windows SDK. (2012) http://msdn.microsoft.com/en-us/windows/hardware/hh852360.aspx(Last accessed, June 2012)
7. Microsoft: SAL annotations. (2012) http://msdn.microsoft.com/en-us/library/ms235402(v=vs.80).aspx Last accessed June 2012
8. Hatada M et al. (2011) Datasets for anti-malware research MWS 2011 Datasets. In: Computer security symposium (CSS2011) Japanese
9. McAfee Inc.: W32/Sdbot.worm. (2009) http://vil.nai.com/vil/content/v_100454.htm, Last accessed, June 2012
10. Symantec Corporation: Backdoor. Sdbot technical details | Symantec. http://www.symantec.com/en/us/security_response/writeup.jsp?docid=2002-051312-3628-99&tabid=2 Last accessed June 2012
11. Trend Micro Incorporated.: PALEVO worm leads to info theft, DDoS attacks | Trend micro threat encyclopedia. (2012) http://about-threats.trendmicro.com/RelatedThreats.aspx?name=PALEVO+Worm+Leads+to+Info+Theft%2C+DDoS+attacks Last accessed June 2012

12. McAfee Inc.: W32/Palevo!4D58C671EE49 - Malware - McAfee labs threat center. (2012) http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=561341 Last accessed, June 2012
13. Sophos Ltd.: 49–2010 - Threat spotlight archive - Threat spotlight - Security news and trends - Sophos. (2012) http://www.sophos.com/en-us/security-news-trends/threat-spotlight/threat-spotlight-archive/2010/49.aspx#f0e736f5-9b72-45c4-a6ec-4cd827fce17a Last accessed Dec 2012
14. McAfee Inc.: W32/Palevo.gen.b!737FE99CE9DB - Malware - McAfee labs threat center. (2012) http://www.mcafee.com/threat-intelligence/malware/default.aspx?id=995696 Last accessed, June 2012
15. Microsoft corporation.: Encyclopedia entry: Virus:Win32/Polip.A - Learn more about malware - Microsoft malware protection center. (2012) http://www.microsoft.com/security/portal/Threat/Encyclopedia/Entry.aspx?Name=Virus%3AWin32%2FPolip.A Last accessed June 2012
16. Symantec Corporation: W32.Polip technical details | symantec. (2012) http://www.symantec.com/security_response/writeup.jsp?docid=2006-042309-1842-99&tabid=2 Last accessed June 2012
17. Olly advanced. (2007) http://www.openrce.org/downloads/details/241/Olly_Advanced
18. PhantOm (2009) - Collaborative RCE tool library. http://www.woodmann.com/collaborative/tools/index.php/PhantOm
19. Ollydbg (2010) v1 10 http://www.ollydbg.de/
20. Vasudevan A, Yerraballi R (2005) Stealth breakpoints. In: Computer security applications conference, 21st Annual, pp 10–392
21. Bayer U et al. (2006) TTAnalyze: a tool for analyzing malware. In: 5th European institute for computer antivirus research (EICAR 2006) Annual conference
22. Anubis (2010) analyzing unknown binaries. http://anubis.iseclab.org/
23. Mandl T et al (2009) Anubis - analyzing unknown binaries the automatic way. Virus bulletin conference. Geneva, Switzerland
24. Bellard F, Qemu, (2005) A fast and portable dynamic translator. Proceedings of the annual conference on USENIX Annual technical conference, USENIX association, Anaheim, CA, pp 41–41
25. Anh QN, Suzaki K (2010) Virt-ice: next generation debugger for malware analysis. Black Hat USA
26. Dinaburg A et al. (2008) Ether: malware analysis via hardware virtualization extensions. In: Proceedings of the 15th ACM conference on computer and communications security, ACM, Alexandria, Virginia, USA pp 51–62
27. Barham P et al (2003) Xen and the art of virtualization. In: Proceedings of the nineteenth ACM symposium on operating systems principles, ACM, Bolton Landing, NY, pp 164–177
28. Microsoft: NtCreateFile function (Windows). (2012) http://msdn.microsoft.com/en-us/library/bb432380.aspx Last accessed June 2012

Chapter 22
Syntatic and Semantic Taxonomy of Preferential Voting Methods

Sung-Hyuk Cha and Yoo Jung An

Abstract Preferential voting where the voters rank candidates in order of preference plays an important role in many decision making problems and have been studied intensively. Yet there are too many variations and many popular methods are promulgated differently in different regions. Hence, some iconic conventional methods are reviewed for syntactic patterns and categorized. A nomenclature for these voting methods is suggested to reveal their syntactic patterns. Over a thousand of voting methods are devised from the conventional procedural patterns. Over 60 representative voting methods are used to reveal their semantic relationship in the form of hierarchical clustering tree. All preferential voting methods perform significantly different from the simplest plurality method.

Keywords Decision · Hierarchical clustering · Nomenclature · Preferential voting · Taxonomy · Voting

22.1 Introduction

Consensus of a group plays an important role in decision making such as elections [1–3] and combining multiple classifiers [4]. It is essential in most democratic societies and has received great attention in artificial intelligence and computer science communities as well [4–6].

S.-H. Cha (✉)
Computer Science Department, Pace University, New York, NY 10038, USA
e-mail: scha@pace.edu

Y. J. An
Division of Engineering Technologies and Computer Sciences, Essex County College,
Newark, NJ 07102, USA
e-mail: yan@essex.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_22,

Consider an ordered set of four candidates, C = {'A', 'B', 'C', 'D'} and they received the corresponding votes, V = {11, 6, 7, 6}. The notations in Table 22.1 shall be used throughout the rest of this chapter. The most widely used and simplest voting method is called the '*plurality*', i.e. the winner is one who has the most votes as defined in Eq. (22.1). The majority voting method in Eq. (22.2) is the same as the plurality but rejects if the winner does not receive more than half votes.

$$plurality(V) = \arg\max_{x \in C}(V_x) \tag{22.1}$$

$$majority(V) = \begin{cases} \arg\max\limits_{x \in C}(V_x) & \text{if } \max(V) > \frac{m}{2} \\ void & \text{otherwise} \end{cases} \tag{22.2}$$

Due to flaws in the simple top choice voting system, the *preferential voting system* in which the voter ranks candidates in order of preference has been proposed [1–3, 6–8]. Voters are asked to rank the candidates where omissions and ties are not allowed and quantities are not important but only the strict order matters as exemplified in Table 22.2.

The winner of Table 22.2 case differs depending on voting methods used and there is a zoo of diverse methods. Various voting methods are used in diverse social groups and different regions. There are so many variation and alternatives and thus a comprehensive study is necessary because even names for certain voting methods are fluid and promulgated differently. Also, a nomenclature for voting methods is necessary as even today another new voting method is invented. This chapter extends the earlier survey work [9].

Table 22.1 Basic notations

Notation	Meaning	Example
C	An ordered set of candidates	{'A', 'B', 'C', 'D'}, C_2 = 'B'
V	The ordered corresponding votes	{11, 6, 7, 6} and V_3 = 7
c	The number of candidates	C = |C| = 4
m	The total number of voters	30 in Table 22.2
n	The number of unique preference order ballots	5 in Table 22.2 ($n \le \min(m, c!)$)
$p(i, j)$	The candidate in the ith ballot and jth choice	p(2, 1) = 'A'
$r(i, x)$	The choice rank for the candidate x in the ith ballot	r(4,'A') = 3

Table 22.2 Sample preference ballot

Choice\votes	7	11	1	6	5
1	C	A	D	B	D
2	D	D	B	C	C
3	B	C	C	A	A
4	A	B	A	D	B

The rest of the paper is organized as follows. In Sect. 22.2, various conventional preference voting methods are given to reveal their syntactic similarities. In Sect. 22.3, conventional methods are generalized and other new methods are devised from the existing methods' patterns. In order to provide a better perspective on similarity among different methods, Sect. 22.3 presents the hierarchical cluster tree of over sixty different preference voting methods. Finally, Sect. 22.4 concludes this work.

22.2 Conventional Voting Methods

In this section, a taxonomy of conventional voting methods is examined, i.e., various voting methods are grouped and classified based on syntactic patterns. Numerous popular preferential voting methods are expressed in as generic mathematical forms as possible.

Perhaps, the most common paradigm for many methods is finding the argmax of certain measurable score values for each candidate as given in Eq. (22.3).

$$sb(p) = \begin{cases} \arg\max\limits_{x \in C}(f_s(p,x)) & \text{if unique} \\ \text{void} & \text{otherwise} \end{cases} \tag{22.3}$$

The popular *plurality* method follows this general form in Eq. (22.3) with Eq. (22.4) as its score function: $f_s(p, x) = pl(p, x)$.

$$pl(p,x) = \sum_{p(i,1)=x} v_i = \sum_{i=1}^{n} \begin{cases} v_i & \text{if } p(i,1) = x \\ 0 & \text{otherwise} \end{cases} \tag{22.4}$$

Note that the summation notation with the subscript only shall be used frequently in this chapter and is defined as the summation of only $f(i,x)$ from $i = 1$ to n such that the subscript condition is met as exemplified in Eq. (22.4).

The paradigm Eq. (22.3), also known as the *rank* method, uses the score function with certain weights. If the weights are (1, 0, 0, 0), it is the plurality method. If the weight is (c−1,…1, 0) as in Eq. (22.5), e.g., (3, 2, 1, 0) in our example, then it is called *borda* score and the Eq. (22.3) with Eq. (22.6) with Eq. (22.5) is called *borda* method [1–4] attributing Jean-Charles de Borda [7].

$$w_{r(i,x)} = c - r(i,x) \tag{22.5}$$

$$f_w(p,x,w) = \sum_{i=1}^{n} w_{r(i,x)} v_i \tag{22.6}$$

Similar to the paradigm Eq. (22.3), finding the *argmin* of certain measurable *penalty* values for each candidate as given in Eq. (22.7).

$$nb(p) = \begin{cases} \underset{x \in C}{\arg\min}(fn(p,x)) & \text{if unique} \\ \text{void} & \text{otherwise} \end{cases} \tag{22.7}$$

As opposed to the simplest score function in the plurality method which considers only the pluralities of the top choice, a simplest penalty function would be the plurality of the last choice as in Eq. (22.8) since the candidate who is disliked by least voters would have the lowest penalty value.

$$fn(p,x) = \sum_{p(i,c)=x} v_i \tag{22.8}$$

Suppose we would like to compare two candidates, ‘B’ and ‘D’. The pairwise comparison (Eq. 22.9) and score (Eq. 22.10) functions have been applied to many conventional methods.

$$pairwin(p,x,y) = \begin{cases} x & \text{if } fp(p,x,y) > fp(p,y,x) \\ y & \text{if } fp(p,x,y) < fp(p,y,x) \\ \text{void} & \text{if } fp(p,x,y) = fp(p,y,x) \end{cases} \tag{22.9}$$

$$fp(p,x,y) = \sum_{r(i,x) < r(i,y)} v_i \tag{22.10}$$

For example of Fig. 22.1a, *fp*(*p*,‘B’,‘D’) = 6 which is the fourth ballot and *fp*(*p*,‘D’,‘B’) = 7 + 11 + 1 + 5 = 24 where ‘D’ precedes ‘B’. Hence, the winner is *pairwin*(*p*,‘B’,‘D’) = ‘D’. Let’s denote this pairwise $c \times c$ victory score matrix based on Eq. (22.10) M as given in Fig. 22.1a.

If there exists a candidate which wins all other candidates in pairwise comparison as defined in Eq. (22.11), this winner is called the *Condorcet* winner [1–3] attributed to Marquis de Condorcet [8]. This concept dates back at least to Ramon Llull in the thirteenth century though [3].

$$condorcet(p) = \begin{cases} x & \text{if } \exists x \forall y (pairwin(p,x,y) = x) \\ & \text{where } x, y \in C \;\&\; x \neq y \\ \text{void} & \text{otherwise} \end{cases} \tag{22.11}$$

	A	B	C	D	*Wt*
A	0	16	11	17	44
B	14	0	7	6	27
C	19	23	0	13	55
D	13	24	17	0	54
Lt	46	63	35	36	

(a)

	A	B	C	D	*Wt*
A	0	1	0	1	2
B	0	0	0	0	0
C	1	1	0	0	2
D	0	1	1	0	2
Lt	1	3	1	1	

(b)

Fig. 22.1 Pairwise victory score and winner matrices **a** Pairwise victory score matrix, M, **b** Pairwise winner matrix, M_b

In the example of Table 22.2 and most of cases, there is no Condorcet winner as shown in Fig. 22.1b. The Condorcet concept is often used as a property of other methods, i.e., whether or not a certain method x always selects the Condorcet winner if there exists one.

Since the borda method does not have the Condorcet property, Duncan Black suggested a method which selects the Condorcet winner if there exists or uses the borda method otherwise in [10]. This method was referred to as the *black* method in [1]. The Condorcet concept can be used generically as an ensemble with other methods which do not have the Condorcet property as in Eq. (22.12).

$$cmethod_x(p) = \begin{cases} w = condorcet(p) & \text{if } w \neq void \\ method_x(p) & \text{otherwise} \end{cases} \tag{22.12}$$

The minimum number of pairwise swaps before they become a Condorcet winner is called the *dodgson* method which is an *NP-hard* problem [11]. In [1], however, a simplified version is attributed to the dodgson method which uses Eq. (22.8) in Eq. (22.12). We shall refer it as dodgson-s method.

Another popular and widely used voting concept involves the *two-round* system. First, it selects the top two candidates by a certain way and then uses the pairwise comparison between those as given in Eq. (22.13).

$$sp(p) = pairwin(p, x_1, x_2) \text{ where } (x_1, x_2) = top2(p) \tag{22.13}$$

Albeit any ranking function such as the borda score can be used to select the top two, the simple plurality (Eq. 22.4) is used in Eq. (22.13) and this particular method is known as the *run-off* method.

Similar to the usage of the Condorcet method (Eq. 22.12), the majority method (Eq. 22.2) can be used as an ensemble with other methods as well.

$$mmethod_x(p) = \begin{cases} w = majority(p) & \text{if } w > \frac{m}{2} \\ method_x(p) & \text{otherwise} \end{cases} \tag{22.14}$$

When Eq. (22.13) is used in Eq. (22.14), this ensemble method is called the *contingent* method or runoff method interchangeably in [1, 2]. Note that the runoff (Eq. 22.13) is the same as Eq. (22.14) with Eq. (22.13) when the plurality (Eq. 22.4) and pairwise comparison (Eqs. 22.9–22.10) functions are used to find and compare the top two candidates, respectively. We make distinction here because the results may differ when other score functions and/or pairwise comparison methods are used in the later Sect. 22.3.

Suppose pairwise matches are scheduled with a fixed sequential agenda as depicted in Fig. 22.2. This method is called the *sequential pairwise match* method in [3]. The winner depends on the order in the agenda, *a*. In Fig. 22.2a the alphabetical order, the winner of ‘A’ vs. ‘B’ match will be against ‘C’. And then the winner of that round will be in the final round with ‘D’. This method is defined recursively in Eq. (22.15). Any ascending or descending order of a score or penalty function can be used as the agenda, e.g., the plurality and borda score orders are given in Fig. 22.2b and c, respectively.

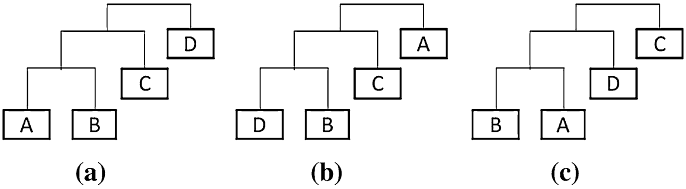


Fig. 22.2 Sequential pairwise match agendas **a** alphabetical order, **b** plurality order, **c** borda score order

$$q(p,a) = \begin{cases} head(a) & \text{if } leng(a) = 1 \\ pairwin(p, head(a), q(p, tail(\mathrm{a}))) & \text{otherwise} \end{cases} \tag{22.15}$$

Suppose that the candidate ‘B’ withdraw from the election. Then the preference ballot table in Table 22.2 is updated to the one in Table 22.3 due to the elimination process of \$p = p − {‘B’}\$. Numerous methods utilize this elimination process.

In 1861, Thomas Hare proposed the *instant runoff* voting method or simply *IRV* which eliminates the candidate with the lowest plurality recursively until there exists a majority winner [3]. The *IRV* is also called *hare* [1], *Cincinnati rule* [2], or *single transferable* [2, 3] method (see [12] for various other names in various countries and regions).

Equation (22.16) is a recursive generic form of the IRV where many possible elimination functions like Eq. (22.17) can be used. The elimination function outputs a set of candidates to be removed. Even if a specific method like the alternative or IRV method requires eliminating a single candidate per step, multiple candidates could be removed as a bulk if there are ties.

$$rmb(p) = \begin{cases} majority(p) & \text{if exists} \\ rmb(p - E) & \text{if } c > 1 \;\&\; E \neq \{\} \\ \quad void & \text{if } c = 0 \text{ or } E = \{\} \end{cases} \tag{22.16}$$

$$E = \arg\min_{x \in C}(pl(p, x)) \tag{22.17}$$

Instead of eliminating the candidate(s) with the fewest first place votes, Clyde Coombs proposed to eliminate those with the most last place votes (Eq. 22.18) [13]. The generic form (Eq. 22.19) with Eq. (22.18) as the elimination function is referred to as the *Coombs* method in [1–3]. While the Hare method may terminate

Table 22.3 p—{‘B’} preference ballot table

Choice\votes	7	11	1	6	5
1	C	A	D	C	D
2	D	D	C	A	C
3	A	C	A	D	A

the elimination when a candidate with majority exists, the Coombs method keeps eliminating candidates until only one remains.

$$E = \arg\max_{x \in C}(pf(p,x)) = \arg\max_{x \in C}\left(\sum_{p(i,c)=x} v_i\right) \tag{22.18}$$

$$rb(p) = \begin{cases} C_1 & \text{if } c = 1 \\ rb(p - E) & \text{if } c > 1 \ \& \ E \neq \{\} \\ void & \text{if } c = 0 \text{ or } E = \{\} \end{cases} \tag{22.19}$$

In 1882, Edward J. Nanson proposed a hybrid method of the generic form (Eq. 22.19) with the borda score where all candidates whose borda scores are below the average are eliminated per recursive step (Eq. 22.20) [14]. Joseph M. Baldwin referred (Eqs. 22.19–22.20) as the *Nanson* method and proposed to eliminate only candidate(s) with the lowest borda score (Eq. 22.21) [15].

$$E = \left\{ x \in C \,\middle|\, bs(x) < \sum_{i=1}^{c} bs(c_i) \middle/ c \right\} \tag{22.20}$$

$$E = \{x \in C | x = \arg\min(bs(*))\} \tag{22.21}$$

The following conventional methods require a $c \times c$ matrix produced by a pairwise comparison function, e.g., M in Fig. 22.1 produced by Eq. (22.10). The *minimax* method, also known as Simpson-Kramer or successive reversal method uses the generic form (Eq. 22.7) with a certain penalty function involving the M_v matrix [2]. The most popular penalty functions include the pairwise opposition (Eq. 22.22), winning votes (Eq. 22.23), and margins (Eq. 22.24).

$$f_{nx\_op}(p,x) = \max(M_v(*,y)) \tag{22.22}$$

$$\begin{aligned} f_{nx\_wv}(p,x) &= \max(Wv) \\ \text{where } Wv(y) &= \begin{cases} M_v(y,z) & \text{if } M_v(z,y) < M_v(y,z) \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{22.23}$$

$$\begin{aligned} f_{nx\_mg}(p,x) &= \max(Mg) \\ \text{where } Mg(y) &= \begin{cases} M_v(y,z) - M_v(z,y) & \text{if } z \neq y \\ -\infty & \text{if } z = y \end{cases} \end{aligned} \tag{22.24}$$

A. H. Copeland suggested a *pairwise aggregation* method [16] which is simply called the Copeland method in [3]. It involves the pairwise winner matrix produced by Eq. (22.25) in Fig. 22.1b.

$$M_b(x,y) = \begin{cases} 1 & \text{if } x = pairwin(p,x,y) \\ 0 & \text{if } y = pairwin(p,x,y) \\ 1/2 & \text{if } void = pairwin(p,x,y) \end{cases} \tag{22.25}$$

Copeland method follows the standard form (Eq. 22.3) with the score function given in Eq. (22.28), i.e., the number of pairwise victories (Eq. 22.26) minus the number of pairwise defeats (Eq. 22.27).

$$S_W(x) = \sum_{y=C_1}^{C_c} M(x,y) \tag{22.26}$$

$$S_L(y) = \sum_{x=C_1}^{C_c} M(x,y) \tag{22.27}$$

$$f_{sx}(x) = S_W(x) - S_L(x) \tag{22.28}$$

22.3 Composition of Voting Methods

This section generalizes the syntactic patterns of conventional voting methods and provides a nomenclature to compose new voting methods. Over a thousand of conventional and composed voting methods are named as strings using the symbols in Table 22.4 and enumerated in Table 22.5.

First, the conventional methods can be categorized into whether it requires any pairwise comparison and two or more rounds as shown in Fig. 22.3. If it does not, the single symbol ‘*b*’ is used and if so, the pair symbol ‘*p*’ or ‘*q*’ shall be used. The generic form (Eq. 22.3) can be named $sb(f_s)$ which uses a certain score function to select the single winner. Instead of the linear weight in the borda score (Eq. 22.6), five other score weights, in which the closer to the top choice, the higher weights it gets, are given in Table 22.6. e.g., the triangular number weight formula is given in Eq. (22.29).

$$w_{r(i,x)} = T_{r(i,x)} = \frac{(c - r(i,x))(c - r(i,x)) + 1)}{2} \tag{22.29}$$

Table 22.4 Symbols in nomenclature in preferential voting methods

	Meaning	Arguments / usages
b	Single	Argmax/ argmin/ leave one, etc
s	Sort by descending order	Requires a score function, f_s
n	Sort by ascending order	Requires a penalty function, f_n
c	Condorcet	Prefix Eq. (22.12)
m	Majority	Prefix Eq. (22.14) / postfix for recursion
p	Pairwise comparison	f_p Eq. (22.9–22.10), Eq. (22.31), Eq. (22.32), Eq. (22.33)
q	Sequential pairwise	Ordering (f_s, f_{sx}, f_n, or f_{nx}) and f_p
r	Recursive elimination	f_e Eq. (22.17), Eq. (22.18), Eq. (22.20), Eq. (22.21)
x	Matrix	f_p to make matrix and (f_{sx} or f_{nx}) to make a vector

Table 22.5 1,001 conventional and composed preferential voting methods

Generic	Conventional	Composed	Count

$sb(f_s)$ (3)	Plurality = $sb(fs_4)$,
	Borda = $sb(fs_1)$
		$sb(fs_2)$, $sb(fs_3)$, $sb(fs_5)$, $sb(fs_6)$	6

$msb(f_s)$ (14)		$msb(fs_1)$, …, $msb(fs_6)$	6
$csb(f_s)$ (12)	Black = $csb(fs_1)$	$csb(fs_2)$, …, $csb(fs_6)$	6
$nb(f_n)$ (7)		$nb(fn_1)$, …, $nb(fn_6)$	6
$mnb(f_n)$		$mnb(fn_1)$, …, $mnb(fn_6)$	6
$cnb(f_n)$	Dodgson-s = $cnb(fn_4)$	$cnb(fn_1)$, …, $cnb(fn_6)$	6
$q(\text{-},f_p)$ (15)	Seq_pair = $q(-, fp_1)$	$q(-, fp_2)$, …, $q(-, fp_7)$	7
$sq(f_s,f_p)$ (15)		$sq(fs_1, fp_1)$, …, $sq(fs_6, fp_7)$	$6 \times 7 = 42$
$nq(f_n,f_p)$ (15)		$nq(fn_1, fp_1)$, …, $nq(fn_6, fp_7)$	$6 \times 7 = 42$
$sp(f_s,f_p)$ (13)	Runoff = $sp(fs_4, fp_1)$	$sp(fs_1, fp_1)$, …, $sp(fs_6, fp_7)$	$6 \times 7 = 42$
$msp(f_s,f_p)$	Contingent = $msp(fs_4, fp_1)$	$msp(fs_1, fp_1)$, …, $msp(fs_6, fp_7)$	$6 \times 7 = 42$
$np(f_n,f_p)$ (31)		$np(fn_1, fp_1)$, …, $np(fn_6, fp_7)$	$6 \times 7 = 42$
$mnp(f_n,f_p)$		$mnp(fn_1, fp_1)$, …, $mnp(fn_6, fp_7)$	$6 \times 7 = 42$
$rbs(f_s)$ (19)	Baldwin = $rbs(fs_1)$	$rbs(fs_2)$, …,$rbs(fs_6)$	6
$rbs*(f_s)$ (19)	Nanson = $rbs*(fs_1)$	$rbs*(fs_2)$, …,$rbs*(fs_6)$	6
$rmbs(f_s)$ (16)	Hare = $rmbs(fs_4)$	$rmbs(fs_1)$, …,$rmbs(fs_6)$	6
$rmbs*(f_s)$ (16)		$rmbs*(fs_1)$, …,$rmbs*(fs_6)$	6
$rbn(f_n)$ (19)	Coombs = $rbn(fn_4)$,	$rbn(fn_1)$,…,$rbn(fn_6)$	6
$rbn*(f_n)$ (19)		$rbn*(fn_1)$, …, $rbn*(fn_6)$	6
$rmbn(f_n)$ (16)		$rmbn(fn_1)$, …,$rmbn(fn_6)$	6
$rmbn*(f_n)$ (16)		$rmbn*(fn_1)$, …, $rmbn*(fn_6)$	6
$rps(f_s)$ (32)		$rps(fs_1)$, …,$rps(fs_6)$	$6 \times 7 = 42$
$rps*(f_s)$ (32)		$rps*(fs_1)$, …,$rps*(fs_6)$	$6 \times 7 = 42$
$rmps(f_s)$ (33)		$rmps(fs_1)$, …,$rmps(fs_6)$	$6 \times 7 = 42$
$rmps*(f_s)$ (33)		$rmps^*(fs_1)$, …,$rmps^*(fs_6)$	$6 \times 7 = 42$
$rpn(f_n)$ (32)		$rpn(fn_1)$,…,$rpn(fn_6)$	$6 \times 7 = 42$
$rpn*(f_n)$ (32)		$rpn*(fn_1)$, …, $rpn^*(fn_6)$	$6 \times 7 = 42$
$rmpn(f_n)$ (33)		$rmpn(fn_1)$, …,$rmpn(fn_6)$	$6 \times 7 = 42$
$rmpn*(f_n)$ (33)		$rmpn*(fn_1)$, …, $rmpn*(fn_6)$	$6 \times 7 = 42$
$sbx(f_p,f_{sx})$	Borda [17] = $sbx(fp_1, fsx_2)$	$sbx(fp_1, fsx_1)$, …, $sbx(fp_7, fsx_5)$	$7 \times 5 = 35$
$nbx(f_p,f_{nx})$	Minimax= $nbx(fp_1, fnx_2)$, $nbx(fp_1,fnx_3)$, $nbx(fp_1,fnx_4)$	$nbx(fp_1, fnx_1)$, …, $nbx(fp_7, fnx_4)$	$7 \times 4 = 28$
$sbxd(f_p, fsx_4)$	Copeland = $sbxd(fp_1, fsx_4)$	$sbxd(fp_1, fsx_4)$, …,$sbxd(fp_7, fsx_4)$	$7 \times 1 = 7$
$rbsx(f_p,f_{sx})$	Baldwin [17] = $rbsx(fp_1,fsx_2)$	$rbsx(fp_1, fsx_1)$, …,$rbsx(fp_7, fsx_5)$	$7 \times 5 =$3 5
$rbsx*(f_p,f_{sx})$	Nanson [17] ≈ $rbsx*(fp_1,fsx_2)$	$rbsx*(fp_1, fsx_1)$, …,$rbsx*(fp_7, fsx_5)$	$7 \times 5 = 35$
$rmbx(f_p,f_{sx})$		$rmbsx(fp_1, fsx_1)$, …,$rmbsx(fp_7, fsx_5)$	$7 \times 5 = 35$
$rmbx*(f_p,f_{sx})$		$rmbsx*(fp_1, fsx_1)$, ..,$rmbsx*(fp_7, fsx_5)$	$7 \times 5 = 35$
$rbnx(f_p,f_{nx})$		$rbnx(fp_1, fnx_1)$, …,$rbnx(fp_7, fnx_4)$	$7 \times 4 = 28$
$rbnx*(f_p,f_{nx})$		$rbnx*(fp_1, fnx_1)$, …,$rbnx*(fp_7, fnx_4)$	$7 \times 4 = 28$
$rmnx(f_p,f_{nx})$		$rmbnx(fp_1, fnx_1)$, ..,$rmbnx(fp_7, fnx_4)$	$7 \times 4 = 28$
$rmnx*(f_p,f_{nx})$		$rmbnx*(fp_1,fnx_1)$, ..,$rmbnx*(fp_7,fnx_4)$	$7 \times 4 = 28$

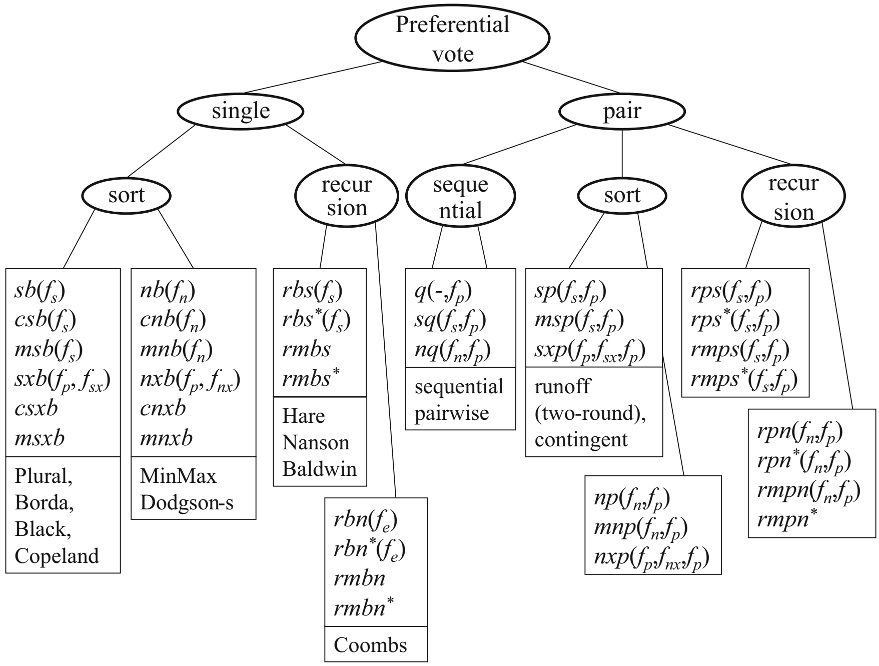

Preferential vote
single
pair
sort
recur sion
seque ntial
sort
recur sion
$sb(f_s)$
$csb(f_s)$
$msb(f_s)$
$sxb(f_p, f_{sx})$
$csxb$
$msxb$
Plural, Borda, Black, Copeland
$nb(f_n)$
$cnb(f_n)$
$mnb(f_n)$
$nxb(f_p, f_{nx})$
$cnxb$
$mnxb$
MinMax Dodgson-s
$rbs(f_s)$
$rbs^*(f_s)$
$rmbs$
$rmbs^*$
Hare Nanson Baldwin
$rbn(f_e)$
$rbn^*(f_e)$
$rmbn$
$rmbn^*$
Coombs
$q(-,f_p)$
$sq(f_s,f_p)$
$nq(f_n,f_p)$
sequential pairwise
$sp(f_s,f_p)$
$msp(f_s,f_p)$
$sxp(f_p,f_{sx},f_p)$
runoff (two-round), contingent
$np(f_n,f_p)$
$mnp(f_n,f_p)$
$nxp(f_p,f_{nx},f_p)$
$rps(f_s,f_p)$
$rps^*(f_s,f_p)$
$rmps(f_s,f_p)$
$rmps^*(f_s,f_p)$
$rpn(f_n,f_p)$
$rpn^*(f_n,f_p)$
$rmpn(f_n,f_p)$
$rmpn^*$

Fig. 22.3 Categorization tree of preference voting methods

Table 22.6 Six score and six penalty functions to be used in Eq. (22.5)

	Score	Meaning.		Penalty	Meaning.
fs_1	<3, 2, 1, 0>	Linear score (Borda) Eq. (22.6)	fn_1	<0, 1, 2, 3>	Linear Eq. (22.30)
fs_2	<6, 3, 1, 0>	Triangular score Eq. (22.29)	fn_2	<0, 1, 3, 6>	Triangular
fs_3	<9, 4, 1, 0>	Quadratic score	fn_3	<0, 1, 4, 9>	Quadratic
fs_4	<1, 0, 0, 0>	Plurality Eq. (22.4)	fn_4	<0, 0, 0, 1>	Least dislike Eq. (22.8)
fs_5	<1, 1, 0, 0>	Dual approval	fn_5	<0, 0, 1, 1>	Dual dislike
fs_6	<1, 1, 1, 0>	Triple approval	fn_6	<0, 1, 1, 1>	Triple dislike

The concepts of Condorcet and Majority are used as combination with other methods such as black and dodgson-s. The symbol ‘*c*’ and ‘*m*’ are used as a prefix for Eqs. (22.12) and (22.14), respectively. The black method can be stated as $csb(fs_1)$. The prefix, ‘*c*’ is redundant for those methods which have the Condorcet property, e.g., sequential pairwise method, Copeland, etc. The prefix, ‘*m*’ can be also redundant, e.g., plurality = $sb(fs_4)$ = $msb(fs_4)$. The generic form (Eq. 22.7) which requires a certain penalty function, f_n is represented as $nb(f_n)$. The dodgson-s method is $cnb(fn_4)$. If the penalty function (Eq. 22.30) which is the reverse borda penalty, is applied to Eq. (22.7), we name it as $nb(fn_1)$. Six different penalty functions are given in Table 22.6.

Table 22.7 Seven pairwise score functions

Name	Winning positional	Positional difference
Constant <1, 1, 1, 1>	$fp_1(p,x,y) = \sum_{r(i,x)<r(i,y)} v(i)$	
Linear <3, 2, 1, 0>	$fp_2 = \sum_{r(i,x)<r(i,y)} (c - r(i,x))v(i)$	$fp_5 = \sum_{r(i,x)<r(i,y)} \lvert r(i,y) - r(i,x)\rvert v(i)$
Triangular <6, 3, 1, 0>	$fp_3 = \sum_{r(i,x)<r(i,y)} T_{r(i,x)}v(i)$	$fp_6 = \sum_{r(i,x)<r(i,y)} \lvert T_{r(i,x)} - T_{r(i,y)}\rvert v(i)$
Quadratic <9, 4, 1, 0>	$fp_4 = \sum_{r(i,x)<r(i,y)} Q_{r(i,x)}v(i)$	$fp_7 = \sum_{r(i,x)<r(i,y)} \lvert Q_{r(i,x)} - Q_{r(i,y)}\rvert v(i)$

$$pf_b(p,x) = \sum_{i=1}^{n} r(i,x)v(i) \tag{22.30}$$

The symbol '*q*' denotes the *sequential pairwise voting* method (Eq. 22.15) where the order agenda and a pairwise score function, *fp* are required. For example, *q*(<A, B, C, D>, fp_1) has the alphabetic agenda and uses (Eq. 22.10) as the pairwise score function in Fig. 22.2a. The first argument indicates the score or penalty function, e.g., $sq(fs_4, fp_1)$ and $nq(fs_1, fp_1)$ use the plurality and borda score as agenda as depicted in Fig. 22.2b and c, respectively. The ascending or descending order from any score or penalty function can serve as the agenda.

The second argument is the pairwise score function like Eq. (22.10). Seven different pairwise score functions are given in Table 22.7. Pairwise score functions like $fp_2 \sim fp_4$ take the winning position into account as scores just like the borda concept. Pairwise score functions like $fp_5 \sim fp_7$ take the difference between winning and losing positions into account as scores.

These new pairwin functions in Table 22.7 can be applied to not only sequential pairwise voting, but also two round systems and IRV instead of the conventional function (Eq. 22.10). The conventional two round system (Eq. 22.13) can be expressed $sp(fs_4, fp_1)$ or $msp(fs_4, fp_1)$. The two arguments are the score function to order the candidates and the pairwise score function to compare the top two candidates. Hence, 42 different methods (6 score × 7 pairwise score functions) are possible for *sp*(*fs*, *fp*) and *msp*(*fs*, *fp*).

Similarly penalty functions in Table 22.6 can be used to find the lowest two candidates and then any pairwise score function can be applied (Eq. 22.31).

$$np(p) = pairwin(p, x_1, x_2) \text{ where } (x_1, x_2) = bottom2(p) \tag{22.31}$$

Next, the symbol '*r*' stands for the recursive elimination where the IRV family methods can be represented. '*rb*' (Eq. 22.19) means keeping eliminating recursively until one single winner remains. '*rbs*' and '*rbn*' mean eliminating the candidate with the minimum score value and the maximum penalty value, respectively, e.g., Coombs = $rbn(fn_4)$ and Baldwin = $rbs(fs_1)$. Any score or penalty functions can be used to find the argmin or argmax for elimination. The star symbol is appended to indicate the method which eliminates all candidates

Table 22.8 Symbols in nomenclature in preferential voting methods

	Score		Penalty	
Single	*rbs*	$E = \{x \in C \mid x = \arg\min(fs_x(^*))\}$	*rbn*	$x = \arg\max(fn_x(^*))$
Mean	*rbs**	$E = \{x \in C \mid fs_x(x) < \sum fs_x(^*)/c\}$	*rbn**	$fn_x(x) > \sum fn_x(*)/c$

below average score value or above average penalty value as given in Table 22.8, e.g., Nanson = *rbs**(fs_1).

The '*m*' symbol is used after '*r*' (Eq. 22.16), e.g., Hare = *rmbs*(fs_4). Note that *rmbs*(*fs*) ≠ *mrbs*(*fs*). In *mrbs*(*fs*), the majority is checked only once at the beginning as in Eq. (22.14) and then *rbs*(*fs*) is executed.

The concept of the runoff (Eq. 22.13) can be used with the recursive elimination. Candidates can be eliminated until two candidates remain as in Eq. (22.32) or Eq. (22.33) instead of a single candidate in Eq. (22.19) or Eq. (22.16). The symbol '*p*' replaces '*b*': *rps*, *rps**, *rmps*, *rmps**, *rpn*, *rpn**, *rmpn*, and *rmpn**.

$$rp(p) = \begin{cases} C_1 & \text{if } c = 1 \\ pairwin(C_1, C_2) & \text{if } c = 2 \\ rp(p - E) & \text{if } c > 1 \& E \neq \{\} \\ void & \text{if } c = 0\, or\, E = \{\} \end{cases} \tag{22.32}$$

$$rmp(p) = \begin{cases} w = majority(p) & \text{if exists} \\ pairwin(C_1, C_2) & \text{if } c = 2 \\ rmp(p - E) & \text{if } c > 2 \& E \neq \{\} \\ void & \text{if } c = 0 \text{ or } E = \{\} \end{cases} \tag{22.33}$$

Finally, the symbol '*x*' stands for the $c \times c$ matrix such as the pairwise victory score matrix in Fig. 22.1a. The matrix score or penalty functions in Table 22.9 can be applied to compose *sbx*(*fp*, *fsx*) or *nbx*(*fp*, *fnx*) where the first argument is the pairwise score function to produce the matrix and the second argument is the matrix score or penalty function. For example, the *minimax* methods can be expressed as *nbx*(fp_1, fnx_2), *nbx*(fp_1, fnx_3) and *nbx*(fp_1, fnx_4). Note that min(*M*(*x*,*)) always returns a vector of zeros since the diagonal is all zeros. Hence, let $\min^{\emptyset}$(*M*(*x*,*)) be the minimum value for each row excluding the diagonal value.

Let '*xd*' denote the pairwise score function which makes the dichotomized matrix (Eq. 22.25) in Fig. 22.1b Then, *copeland* method can be expressed as *sbxd*(fp_1, fsx_4). Only fsx_4 seems to be applicable in our experiments.

The Borta method is defined differently without any reference in [17] and can be expressed as *sbx*(fp_1, fsx_2). Let's denote it as bs_2. For the Baldwin and Nanson methods, bs_2 is used instead of *borda score* in [17]. The nanson method in [17] used Eq. (22.34) instead of Eq. (22.20).

$$E = \{x \mid x \in C \& bs_2(x) < 0\} \tag{22.34}$$

The Baldwin in [17] can be expressed as *rbsx*(fp_1, fsx_2) and the Nanson in [17] can be approximated by the *rbsx**(fp_1, fsx_2) method.

Table 22.9 Five matrix score functions and four matrix penalty functions

Matrix score functions					
fsx_1	$\sum_{y=C_1}^{C_c} M(x,y)$ Eq. (26)	fsx_2	$\sum\limits_{y=C_1}^{C_c} M(x,y) - \sum\limits_{x=C_1}^{C_c} M(x,y)$ Eq. (28)		
fsx_3	$\max(M(x,^*))$	fsx_4	$\max(M(x,^*))-$ $\max(M(^*,y))$	fsx_5	$\min^{\emptyset}(M(x,^*))$
Matrix penalty functions					
fnx_1	$\sum_{x=C_1}^{C_c} M(x,y)$ Eq. (27)	fnx_2	$\max(M(^*,y))$ Eq. (22)		
fnx_3	$\max(Wv), \text{ where } Wv(y,z) = \begin{cases} M(y,z) & \text{if } M(z,y) < M(y,z) \\ 0 & \text{otherwise} \end{cases}$ Eq. (23)				
fnx_4	$\max(Mg), \text{ where } Mg(y,z) = \begin{cases} M(y,z) - M(z,y) & \text{if } z \neq y \\ -\infty & \text{if } z = y \end{cases}$ Eq. (24)				

22.4 Hierarchical Clustering

Hitherward, the focus is moved from the syntactic patterns to the semantic similarity between voting methods. So as to assess how much similar or different voting methods are, the following experiments were conducted using the *hierarchical cluster analysis* (see [5] for the process description). For m = 100 voters and c = 4 candidates, (nt = 100) number of preference ballot test cases are randomly generated. The distance between two voting methods is the probability of the mismatches in (22.35).

$$d(vm_x, vm_y) = \sum_{i=1}^{nt} \begin{cases} 0 & \text{if } vm_x(i) = vm_y(i) \\ 1 & \text{if } vm_x(i) \neq vm_y(i) \end{cases} \Big/ nt \tag{22.35}$$

If two methods are identical, the distance is zero and if they do not agree on many cases, the distance is high.

The hierarchical cluster tree in Fig. 22.4 reveals the similarities among various voting methods. Prior to rigorous mathematical proof, this *dendrogram* provides the intuitive identity, similarity, correctness, etc. The plurality *sb*(*fs*$_4$), dual approval *sb*(*fs*$_5$), and dodgson-s *cnb*(*fn*$_4$) methods are outliers, i.e., quite different from most other preference voting methods.

In this experiment, the *Borda sb*(*fs*$_1$), *nb*(*fn*$_1$), *q*(*<abcd>*, *fp*$_5$), *sq*(*fs*$_4$, *fp*$_5$), *np*(*fn*$_1$, *fp*$_5$), *Borda* [17] *sbx*(*fp*$_1$, *fsx*$_2$), *nbx*(*fp*$_1$, *fnx*$_1$), *nbx*(*fp*$_5$, *fnx*$_3$), *sbxd*(*fp*$_5$, *fsx*$_2$) turned out to be identical. The *Baldwin rbs*(*fs*$_1$) is identical to *rmbs*(*fs*$_1$) but not to the Baldwin [17] *rbsx*(*fp*$_1$, *fsx*$_2$). Nanson *rbs**(*fs*$_1$) is experimentally identical to *rbsx*(*fp*$_1$, *fsx*$_2$) and *rmbs**(*fs*$_1$). Another identical group is the minimax group: {*snx*(*fp*$_1$, *fnx*$_2$), *snx*(*fp*$_1$, *fnx*$_3$), *snx*(*fp*$_1$, *fnx*$_4$)}. The prefix *m* is sometimes redundant, e.g., *msp*(*fs*$_4$, *fp*$_1$) and *sp*(*fs*$_4$, *fp*$_1$) are identical. Other experimental identical groups

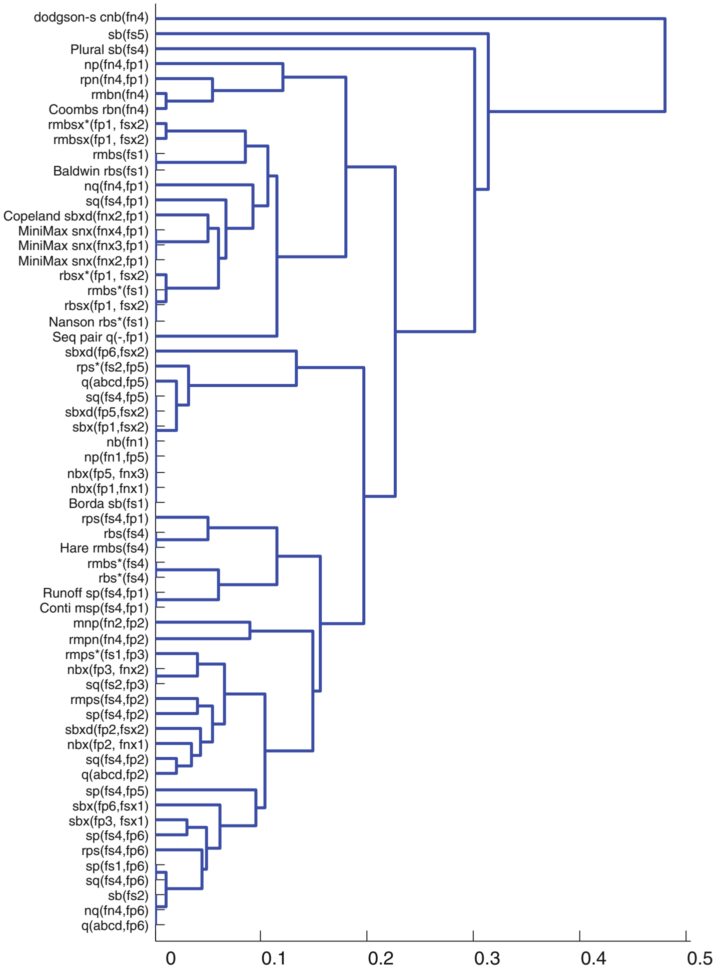


Fig. 22.4 Hierarchical clustering of 51 conventional and devised voting methods

include {*Hare rmbs*(*fs*$_4$), *rbs*(*fs*$_4$)}, {*Coombs rbn*(*fn*$_4$), *rmbn*(*fn*$_4$)}, {*sb*(*fs*$_2$), *sp*(*fs*$_1$, *fp*$_6$)}, {*rbs**(*fs*$_4$), *rmbs**(*fs*$_4$)}.

22.5 Conclusion and Future Work

This chapter revealed the syntactic and semantic relationships among preferential voting methods. Albeit it needs community based consensus and further refinement, an initial naming convention is suggested using the finite number of symbols. Over a thousand syntactically different preferential voting methods are enumerated in Table 22.5. 61 representative voting methods are used to reveal their semantic relationship in the hierarchical clustering tree.

This chapter reviewed several popular and rudimental methods but there are numerous other voting methods are in use. Further comprehensive survey is necessary as a future work. Semantic relationship may vary depending on the number of candidates. Further experiments are necessary for different number of candidates.

References

1. Samuel M III (1988) Making multicandidate elections more democratic. Princeton University Press, Princeton
2. Levin J, Nalebuff B (1995) An introduction to vote-counting schemes. J Econ Perspect 9(I):3–26 Winter
3. Taylor AD, Pacelli AM (1995) Mathematics and politics: strategy, voting, power and proof. Springer, Berlin
4. Ho TK, Hull JJ, Srihari SN (1984) Decision combination in multiple classifier systems. IEEE Trans J PAMI 16(I):66–75
5. Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, London
6. Faliszewski P, Hemaspaandra E, Hemaspaandra LA, Rothe J (2009) Llull and Copeland voting computationally resist bribery and constructive control. J Artif Intell Res 35:275–341
7. de Borda JC (1781) Mémoire sur les Élections au Scrutin. Histoire del l'Académie Royale des Sciences, Paris
8. de Condorcet M (1785) Essay on the application of mathematics to the theory of decision-making. Paris
9. Cha S, An YJ (2012) Taxonomy and nomenclature of preferential voting methods, Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science, WCECS 2012, San Francisco, USA, pp 173–178, 24–26 Oct 2012
10. Black D (1953) The theory of committees and elections. Cambridge University Press, London
11. Bartholdi J III, Tovey CA, Trick MA (1989) Voting schemes for which it can be difficult to tell who won the election. Soc Choice Welfare 6(2):157–165
12. Reilly B (2003) The global spread of preferential voting: Australian institutional imperialism. Aust J Political Sci 39(2):253–266
13. Coombs C (1964) A theory of data. Wiley, New York
14. Nanson EJ (1882) Methods of election. Trans Proc R Soc Victoria 19:197–240
15. Baldwin JM (1926) The technique of the Nanson preferential majority system of election. Proc R Soc Victoria 39:42–52
16. Copeland AH (1951) A 'reasonable' social welfare function. Presented at the seminar on mathematics in social sciences, University of Michigan
17. LeGrand R (2012) Descriptions of ranked-ballot voting methods. http://www.cs.wustl.edu/~legrand/rbvote/desc.html as of June 2012

Chapter 23
High Assurance Enterprise Scaling Issues

William R. Simpson and Coimbatore Chandersekaran

Abstract Many Organizations are moving to web-based approaches to computing. As the threat evolves to higher levels of sophistication, many governmental and commercial organizations are also moving toward high assurance. This chapter describes an approach that uses strong bi-lateral end-to-end authentication with end-point encryption and with SAML-based authorization using OASIS Security Standards. This service-based approach offers many of the advantages of the cloud-based approaches. Cloud-based approaches allow for more agile scale-up, while maintaining a low marginal cost of accommodating increased users. However, many of the applications require high assurance, attribution, formal access control processes, and a wide range of threat mitigation procedures for many of the industries (banking, credit, content distribution, etc.) that are considering conversion to cloud computing environments. Current implementations of cloud services do not meet these high assurance requirements. This high assurance requirement presents many challenges to normal computing and some rather precise requirements that have developed from high assurance issues for web service applications. Gearing up for a large number of users is often difficult without security issues. The most difficult part of scaling up to higher user levels is the maintenance of the security paradigms that provide mitigation of these generic and specific threats. Several issues relating to large scale use that are specific to high assurance and their solutions are discussed at length.

Keywords Assurance · Attribution · Authentication · Authorization · Cloud · Hypervisor · Security · PKI · SAML · Virtualization

W. R. Simpson (✉) · C. Chandersekaran
Institute for Defense Analyses, 4850 Mark Center Drive, Alexandria, VA 22311, USA
e-mail: rsimpson@ida.org

C. Chandersekaran
e-mail: cchander@ida.org

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_23,

23.1 Introduction

This paper is based in part on a paper published by WCECS [1]. In certain enterprises, the network is continually under attack. Examples might be:

- Banking industry enterprise such as a clearing house for electronic transactions,
- Defense industry applications,
- Credit card consolidation processes that handle sensitive data, both fiscal and personal,
- Medical with concerns for privacy and statutory requirements,
- Content distributors worried about rights in data, or theft of content.

The attacks have been pervasive and continue to the point that nefarious code may be present, even when regular monitoring and system sweeps clean up readily apparent malware. This omnipresent threat leads to a healthy paranoia of resistance to observation, intercept, and masquerading. Despite this attack environment, the web interface is the best way to provide access to many of the enterprise users. One way to continue operating in this environment is to not only know and vet your users, but also your software and devices.

Even that has limitations when dealing with the voluminous threat environment. Today we regularly construct seamless encrypted communications between machines through SSL or other TLS. These do not cover the "last mile" between the machine and the user (or service) on one end, and the machine and the service on the other end. This last mile is particularly important when we assume that malware may exist on either machine, opening the transactions to exploits for eavesdropping, ex-filtration, session high-jacking, data corruption, man-in-the-middle, masquerade, blocking or termination of service, and other nefarious behaviors. Before we examine the challenges of large scale computing systems, let us first examine what high assurance architecture might look like.

23.2 Tenets for High Assurance Computing

This section provides nine tenets that guide decisions in an architectural formulation for high assurance and implementation approaches [2]. These tenets are separate from the "functional requirements" of a specific component (e.g., a name needs to be unique); they relate more to the goals of the solution that guide its implementation.

- The *zeroth* tenet is that the malicious entities can look at all network traffic and send virus software to network assets. In other words, rogue agents (including insider threats) may be present and to the extent possible, we should be able to operate and, in their presence, although this does not exclude their ability to view some activity. Assets are constantly monitored and cleaned; however, new attacks may be successful at any time and nefarious code may be present at any given time.

- The *first* tenet is simplicity. This seems obvious, but it is notable how often this principle is ignored in the quest to design solutions with more and more features. That being said, there is a level of complexity that must be handled for security purposes and implementations should not overly simplify the problem for simplicity's sake.
- The *second* tenet, and closely related to the first is extensibility. Any construct we put in place for an enclave should be extensible to the domain and the enterprise, and ultimately to cross-enterprise and coalition. It is undesirable to work a point solution or custom approach for any of these levels.
- The *third* tenet is information hiding. Essentially, information hiding involves only revealing the minimum set of information to the outside world needed for making effective, authorized use of a capability. It also involves implementation and process hiding so that this information cannot be farmed for information or used for mischief.
- The *fourth* tenet is accountability. In this context, accountability means being able to unambiguously identify and track what active entity in the enterprise performed any particular operation (e.g., accessed a file or IP address, invoked a service). Active entities include people, machines, and software process, all of which are named registered and credentialed. By accountability we mean attribution with supporting evidence. Without a delegation model, and detailed logging, it is impossible to establish a chain of custody or do effective forensic analysis to investigate security incidents.
- The *fifth* tenet is minimal detail (to only add detail to the solution to the required level). This combines the principles of simplicity and information hiding, and preserves flexibility of implementation at lower levels. For example, adding too much detail to the access solution while all of the other IA components are still being elaborated may result in wasted work when the solution has to be adapted or retrofitted later.
- The *sixth* tenet is the emphasis on a service-driven rather than a product-driven solution whenever possible. Using services makes possible the flexibility, modularity, and composition of more powerful capabilities. Product-driven solutions tend to be more closely tied to specific vendors and proprietary products. That said, commercial off-the-shelf (COTS) products that are as open as possible will be emphasized and should produce cost efficiencies. Procurement specifications should require functionality and compatibility in lieu of requiring operations in a Microsoft forest environment.
- The *seventh* tenet is that lines of authority should be preserved and IA decisions should be made by policy and/or agreement at the appropriate level.
- The *eighth* tenet is need-to-share as overriding the need-to-know. Health, defense, and finance applications often rely upon and are ineffective without shared information.

23.2.1 Architectural Features

Building an architecture that conforms to these tenets requires specific elements to ensure the tenets are built into systems. In the architecture we espouse, the basic formulation follows a web 2.0 approach and uses Organization for the Advancement of Structured Information Standards (OASIS) standards of security [4]. These elements are listed in the following sections.

23.2.1.1 Naming and Identity

Identity will be established by the requesting agency. All recognized certificate authorities naming scheme must be honored. To avoid collision amongst the schemes, the identity used by all federated exchanges shall be the distinguished name as it appears on the primary credential provided by the certificate authority. The distinguished name must be unique over time and space, which means that retired names are not reused and ambiguities are eliminated. Naming must be applied to all active entities (persons, machines, and software).

23.2.1.2 Credentials

Credentials are an integral part of the federation schema. Each identity (all active entities) requiring access shall be credentialed by a trusted credentialing authority. Further, a Security Token Server (STS) must be used for storing attributes associated with access control. The STS that will be used for generating Security Assertion Markup language (SAML) tokens must also be credentialed (primarily through the same credentialing authority, although others may be entertained).

23.2.1.3 PKI Required: X.509 Certificates

The primary exchange medium for setting up authentication of identities and setting up cryptographic flows is the Public Key Infrastructure (PKI) embodied in an X.509 certificate from a trusted Certificate Authority (CA).

23.2.1.4 Certificate Services

The CA must use known and registered (or in specific cases defined) certificate revocation and currency-checking software. The CA is placed in the root trust store.

23.2.1.5 Bi-Lateral End-to-End Authentication

The requestor will not only authenticate to the service (not the server), but the service will authenticate to the requestor. This two-way authentication avoids a number of threat vulnerabilities. The requestor will initially authenticate to the server and set up a Secure Socket Layer (SSL) connection to begin communication with the service. The primary method of authentication will be through the use of public keys in the X.509 certificate, which can then be used to set up encrypted communications (either by X.509 keys or a generated session key). The preferred method of communication is secure messaging, contained in Simple Object Access Profile (SOAP) envelopes. All messages are encrypted for delivering to the recipient of the message.

23.2.1.6 Authorization Using SAML Packages

All authorizations will be through the use of SAML packages in accordance with the SAML 2.0 specification provided by OASIS [3–8]. The SAML will not be used for authentication, and the SAML is bound to the bi-lateral authentication by a match of the Distinguished Name in both the SAML and the PKI certificate used to authenticate.

23.2.1.7 Registration of the STS

All STS that create and sign SAML packages must be registered. The certificate of the STS will be used to sign SAML tokens, and complete bi-lateral authentication between requestors and the STS.

23.2.1.8 Recognizing STS Signatures

STS signatures will be recognized only for registered STSs (federation and non-federation) and may be repackaged by the local STS when federation registration has been accomplished. Unrecognized signatures will not be honored, access will be denied, and the refusal will be logged as a security relevant event.

23.2.1.9 Recognizing STS Certificates

Local STSs within the enterprise forests will maintain a certificate registration store of all recognized STSs (federated and non-federated). Federated STSs will contain agreed to mappings to facilitate the re-issuance of SAML packages when appropriate.

23.3 High Assurance Architecture Elements

Despite the obvious advantages of cloud computing, the large amount of virtualization and redirection poses a number of problems for high assurance. In order to understand this, let's examine a security flow in a high assurance system.

The basic elements include a user, who initially authenticates to his/her domain using a hardware token and establishes a Virtual Private Network (VPN) session; a Security Token Server (STS), in this case the Identity Provider (IdP); and attribute stores for generating the Security Assertion Markup Language (SAML) token.

The application system consists of a web application (for communication with the user), one or more aggregation services that invoke one or more exposure services and combine their information for return to the web application and the user. As a prerequisite to end-to-end communication an SSL or other suitable TLS is set up between each communication of the machines. The exposure services retrieve information from one or more Authoritative Data Sources (ADSs). Each communication link in Fig. 23.1 will be authenticated end- to-end with the use of public keys in the X.509 certificates provided for each of the active entities. This two-way authentication avoids a number of threat vulnerabilities. The requestor initially authenticates to the service provider. Once the authentication is completed, a TLS/SSL connection is established between the requestor and the service provider, within which a WS-Security package will be sent to the service. The WS-Security package contains a SAML token generated by the Security Token Server (STS) in the requestor domain. The primary method of authentication will be through the use of public keys in the X.509 certificate, which can then be used to set up encrypted communications (either by X.509 keys or a generated session key). Session keys and certificate keys need to be robust and sufficiently protected to prevent malware exploitation. The preferred method of communication is secure messaging using WS Security, contained in SOAP envelopes. The encryption key used is the public key of the target (or a mutually derived session key), ensuring only the target can interpret the communication.

The problem of scale-up and performance is the issue that makes cloud environments so attractive. The cloud will bring on assets as needed and retire them as needed. The trick is to maintain the security paradigm as we scale up.

A traffic cop (load balancer) monitors activity and posts a connection to an available instance. In this case all works out since the new instance has a unique name, end-point, and credentials with which to proceed. All of this, of course needs to be logged in a standard form and parameters passed to make it easy to reconstruct for forensics. We have shown a couple of threats that need mitigation where one eavesdrops on the communication and may actually try to insert himself into the conversation (man-in-the-middle). This highlights the importance of bi-lateral authentication and encrypted communications. The second highlights the need to protect caches and memory spaces.

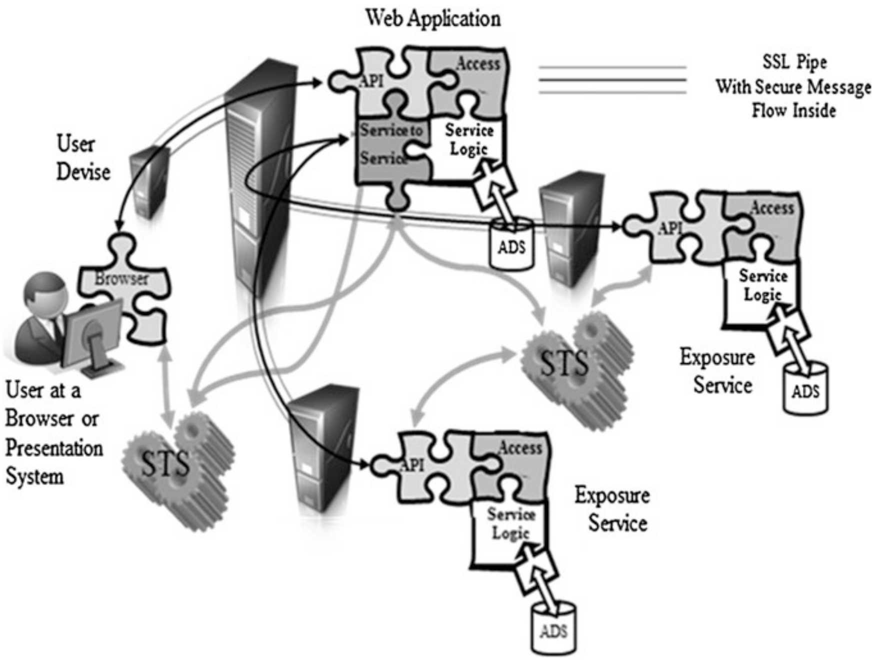


Fig. 23.1 High assurance security flows

23.4 High Assurance Scaling Issues

Among the many scaling issues in a large-scale enterprise, four stand out as related to the high-assurance issues:

- Complex calling sequence (first go get a SAML, then go to service). Example: www.securitytokenserver12.mil/tokenrequest//www.myapplication.mil
- An STS is involved at the initiation of every session.
  - 500,000 users
  - Multiple Services
- Many Services will need load balancing to meet users' needs
- All session are bi-laterally authenticated by PKI and encrypted (TLS)

23.5 Complex Calling Sequence

In order to make the first communication to a home page in the service environment, we will assume a web application [Enterprise Services Homepage (ESHP)] or a device application [Enterprise Application Store (EAPPS)] that can provide the user links to appropriate services. These links are complex in that they must contain the Request for SAML Token (RST), the URI of the IdP, and the URI of the target application. The first link to the ESHP will be provided as a widget on the Enterprise Standard desktop. This ESHP or EAPPS will bilaterally authenticate with the user and consume the user's SAML in order to provide the user an appropriate list of services.

The initial web application will contain all of the information necessary to provide this to the user. To do so, the web application must have access to or contain a registry with the following information:

- An enumerated list of Authorized SAML Producers
  - Includes Public Key (for SAML consumption)
  - Includes URI of the SAML Producer (IdP).
- An enumerated list of web applications
  - Unique name of each application, or Traffic Handling Web Application (see discussion below)
  - The end point URI for accessing the applications or Traffic Handling Web Application (see discussion below)
  - A description of the service
  - The Access Control List (ACL) for the Service
  - The URI of the IdP for the user.

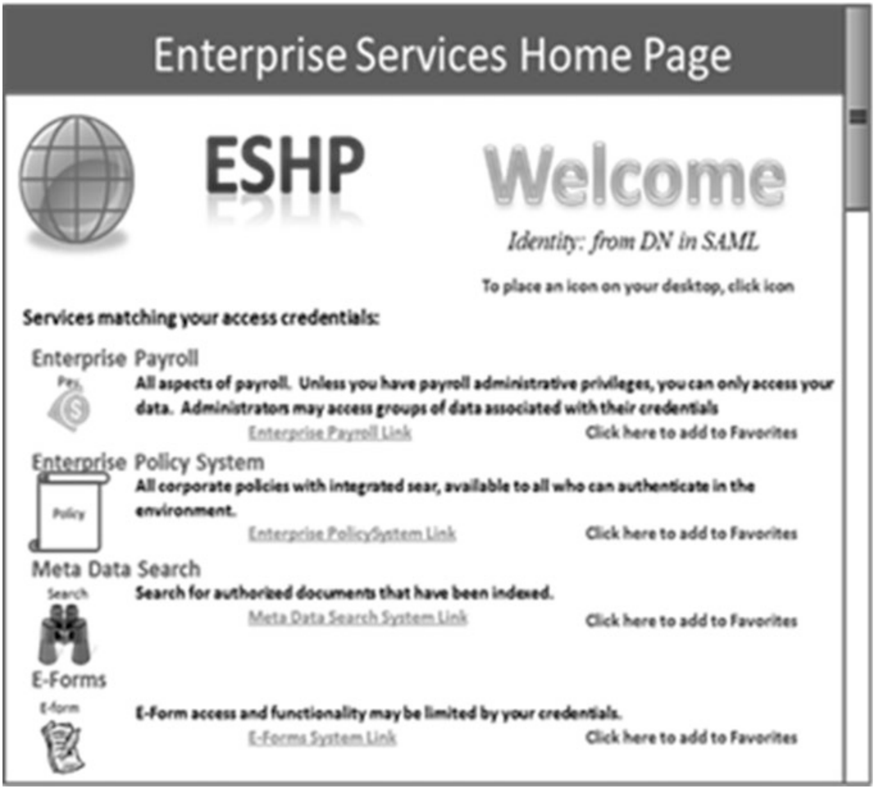
The last element is used to eliminate services for which the user does not have authorization. This is done by matching the unfiltered SAML elements to the ACL in the last bullet. A match must occur (or the ACL is none) in order for the ESHP or EAPPS to provide a link to the service. When the match occurs, the ESHP will add the name, descriptive material, and the complex URI link to the home page.

All registered services are examined and when the list is compiled, the home page is sent to the user. A notional homepage is provided in Fig. 23.2. A notional device display for the EAS is presented in Fig. 23.3. The desktop or device icons must necessarily expire after a time period (say three months) to provide security against changes in status or privilege. When a change occurs, the authorization will fail, but the expiring of the icon will clean up the displays on the desktop or device.

Selecting an icon will invoke a link that will send an RST and a compound URI to the IdP for processing and connect the user with the end-point for bi-lateral authentication with a SAML token at the application. If the SAML is acceptable, a session is started between the user and the web application.

23.6 Scaling the STS

Scale-up to higher levels of users will require a number of different schemes. The most critical, since it involves every request, will be the STS. Example data needed for load-balancing calculations are provided below:



Enterprise Services Home Page

ESHP

Welcome

Identity: from DN in SAML

To place an icon on your desktop, click icon

Services matching your access credentials:

Enterprise Payroll

All aspects of payroll. Unless you have payroll administrative privileges, you can only access your data. Administrators may access groups of data associated with their credentials

Enterprise Payroll Link Click here to add to Favorites

Enterprise Policy System

All corporate policies with integrated sear, available to all who can authenticate in the environment.

Enterprise PolicySystem Link Click here to add to Favorites

Meta Data Search

Search for authorized documents that have been indexed.

Meta Data Search System Link Click here to add to Favorites

E-Forms

E-Form access and functionality may be limited by your credentials.

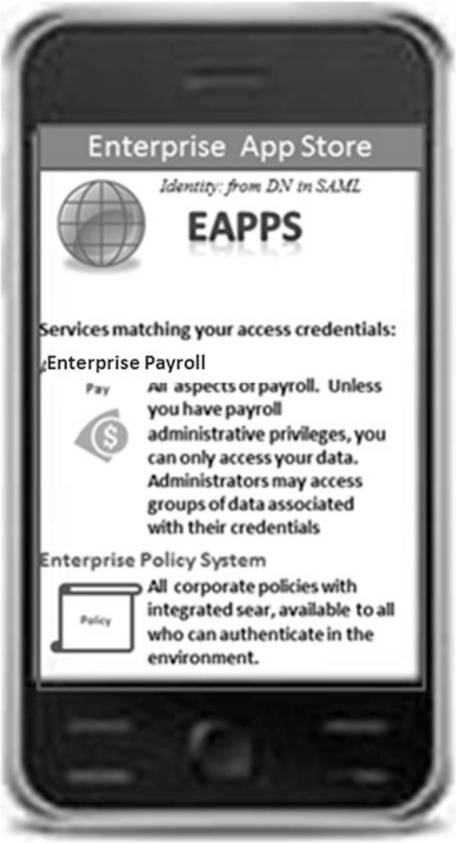
E-Forms System Link Click here to add to Favorites

Fig. 23.2 Enterprise service home page

- Test data are still being developed; however, assume testing indicates 100 token requests can be satisfied in 1 s by the current STS (IdP). Improved versions of the STS or processors may reduce these requirements
- Requirement for 1,667 SAML tokens per second (500,000 users in 5 min)
  - Need 23 STS (IdP)—assumes a 25 % throughput loss in multiple clustering.

Note that the STS is a trusted component and can be load balanced in the traditional manner, as shown in Fig. 23.4.

In this configuration, the clusters share the same naming, PKI credentials and end-point identities. This is the only exception to the requirements for uniquely naming, PKI credentials, and end-points because the STS is the primary trusted component in the system. The need for load balancing of the SP is not anticipated at this time, since it is used for occasional SAML verification or SAML rewriting in the case of federation.



Enterprise App Store

Identity: from DN in SAML

EAPPS

Services matching your access credentials:

Enterprise Payroll

All aspects of payroll. Unless you have payroll administrative privileges, you can only access your data. Administrators may access groups of data associated with their credentials

Enterprise Policy System

All corporate policies with integrated sear, available to all who can authenticate in the environment.

Fig. 23.3 Enterprise application store flows

23.7 Scaling of Services

The second most likely element to need scale-up is the ESHP. This one is a bit trickier. Schemas that essentially extend the thread capabilities of the application may be employed. When the capacities of the thread architecture are exhausted and independent instance of the application need to be set up, the process requires care. The application or service is not a trusted element and is not exempt from the requirements for unique naming, credentialing, and end-points. Each independent instance must have its own name, URI, and credentials. Further, each independent instance must be provisioned for in the attribute stores if it will make further service calls. Fortunately, all web applications and services can be handled in the same manner and the process is not dependent on the number of instances needed to handle the user request. A specific (x, where x corresponds to the web application or web service that needs balancing) Instance Availability Service (IASx) is

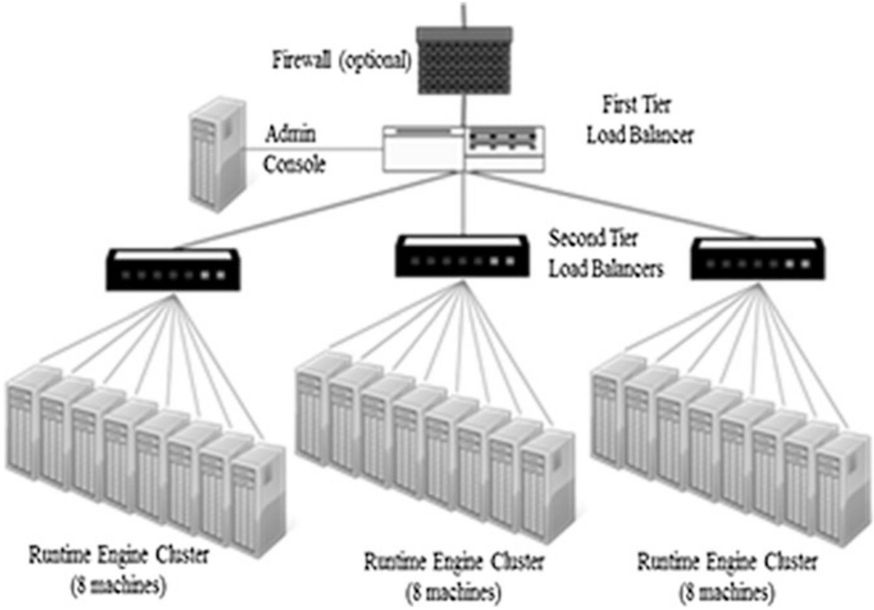


Fig. 23.4 STS load balancing store flows

set up as the end point for a request that needs to be load balanced. Note that this means that the end-point in the ESHP must be changed to the specific IASx, and the end-point in the widget for the ESHP must be changed to the IAS for the ESHP. The IASx needs to have the following information available:

- Number of independent instances of the web application
- Unique name of each independent instance of the web application
- Unique end-point of each independent instance of the web application
- (OPTIONALLY) Usage and load data for each of the independent instances of the web application.

The user goes to the IdP with the address of the IAS, which he can have stored in his favorites, or executes the link on the ESHP (or in the case of ESHP from the desktop widget). The IdP then posts the SAML over HTTPS to the IAS. The IAS doesn't even need to read the SAML (authentication only, or identity-based access control), but would repost the SAML over HTTPS to the independent instance it calculates. It is then completely out of the way, just like the IdP is completely out of the way, and the user is in session with the instance of the web application. This process is shown in Fig. 23.5.

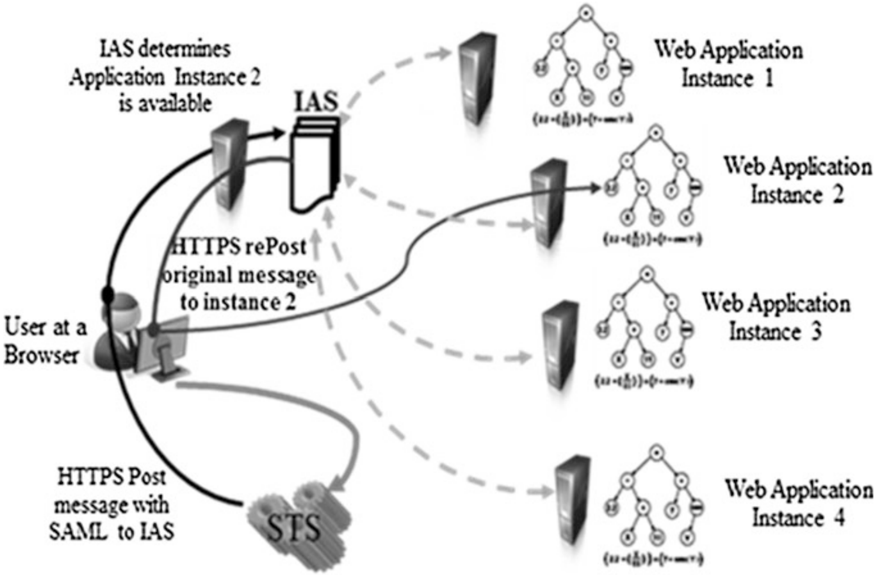

IAS determines Application Instance 2 is available
IAS
Web Application Instance 1
Web Application Instance 2
Web Application Instance 3
Web Application Instance 4
HTTPS rePost original message to instance 2
User at a Browser
HTTPS Post message with SAML to IAS
STS

Fig. 23.5 High assurance load balancing

23.8 Scaling of Services and Maintaining Credentials

Key management is complex and essential. When a new instance is required, it must be built and activated (credentials and properties in the attribute store, as well as end-point assignment). When a current instance is retired, it must be disassembled, and de-activated (credentials and properties in the attribute store, as well as end-point assignment). All of these activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics.

Rules for maintaining high assurance during service load balancing:

1. Shared Identities and credentials break the accountability paradigm.

   - Each independent instance of a machine or service must be uniquely named [9] and provided a PKI Certificate for authentication. The Certificate must be activated while the virtual or real machine is in being, and de-activated when it is not, preventing hijacking of the certificate by nefarious activities.
   - The naming and certificates must be pre-issued and self- certification is not allowed. Each independent instance of a machine or service must have a unique end point. This may take some manipulation through the load balancing process but is required by attribution and accountability. This means that simple re-direct will not work. The one exception is the IdP of the STS, which is trusted software. Extensions of the thread mechanism by assigning resources to the operating system may preserve this functionality. The

individual mechanism for virtualization will determine whether this can be accomplished.

2. Each independent instance of a service must have an account provisioned with appropriate elements in an attribute store. These must be pre-issued and linked to the unique name for each potential independent instance of a service. This is required for SAML token issuance.
3. The importance of cryptography cannot be overstated, and all internal communications as well as external communications should be encrypted to the end point of the communication. Memory and storage should also be encrypted to prevent theft of cached data and security parameters.
4. Private keys must reside in Hardware Storage Modules (HSMs). The security of the Java software key store does not meet high assurance criteria. Stand-up of an independent machine or service must link keys in HSM, and activate credentials pre-assigned to the service.
    - Stand-down of an independent machine or service must de-link keys in HSM, and de-activate credentials pre-assigned to the service.
    - Key management in this environment is a particular concern and a complete management schema including destruction of session keys must be developed.
5. Proxies and re-directs break the end-to-end paradigm.
    - When end points must change, a re-posting of communication is the preferred method. There must be true end-to-end communication with full attribution. This will mean that communication must be re-initiated from client to server when an instance is instantiated, and it must have a unique end point, with unique credentials and cryptography capabilities.
6. All activities must be logged in a standard format with reference values that make it easy to reassemble the chain of events for forensics.

23.9 Summary

We have reviewed the basic approaches to scale up in high assurance computing environments. Virtualization must be very carefully reviewed to ascertain if the security paradigm can be maintained. Extensions of the thread mechanism by assigning resources to the operating system may preserve this functionality. The individual mechanism for virtualization will determine whether this can be accomplished. Notably the extensive use of virtualization and redirection is severe enough that many customers who need high assurance have moved away from the concept of cloud computing. Figure 23.6 provides a summary of how a user addresses an individual web application in a scaled-up system. This processes described in this paper are part of a broad-scale, high-assurance enterprise

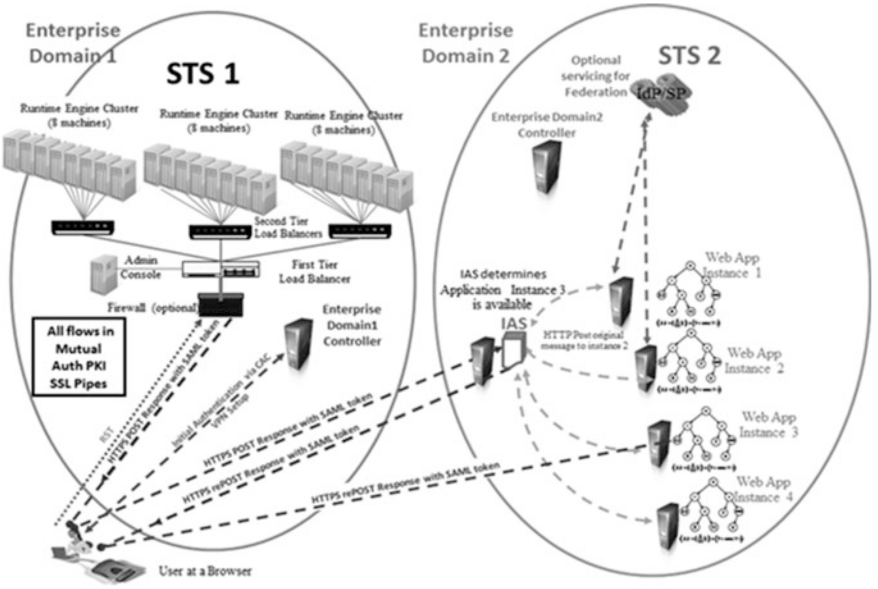

Enterprise Domain 1
STS 1
Runtime Engine Cluster ($ machines)
Runtime Engine Cluster ($ machines)
Runtime Engine Cluster ($ machines)
Second Tier Load Balancers
Admin Console
First Tier Load Balancer
Firewall (optional)
Enterprise Domain1 Controller
All flows in Mutual Auth PKI SSL Pipes
RST
HTTPS POST Response with SAML token
Initial Authentication via CAC VPN Setup
HTTPS POST Response with SAML token
HTTPS rePOST Response with SAML token
HTTPS rePOST Response with SAML token
User at a Browser
Enterprise Domain 2
STS 2
Optional servicing for Federation
IdP/SP
Enterprise Domain2 Controller
IAS determines Application Instance 3 is available
IAS
HTTP Post original message to instance 2
Web App Instance 1
Web App Instance 2
Web App Instance 3
Web App Instance 4

Fig. 23.6 Accessing a web application in a scaled up environment

stand-up, including high-assurance specification, and current implementation. Other aspects of these enterprise processes are described in [10–15].

References

1. Simpson WR (2012) Lecture notes in engineering and computer science. In: Proceedings world congress on engineering and computer science 2012, Enterprise high assurance scale-up, vol 1. San Francisco, pp 54–59
2. Chandersekaran C (2009) Air force information assurance strategy team, air force information assurance enterprise architecture, version 1.70, SAF/XC. (Not available to all)
3. Shibboleth Project (2011) Available at http://shibboleth.internet2.edu/
4. OASIS Identity Federation (2005a) Web service security: scenarios, patterns, and implementation guidance for web services enhancements (WSE) 3.0, Microsoft Corporation
5. OASIS Identity Federation (2005b) WSE 3.0 and WS-reliablemessaging, Microsoft white paper. Available at http://msdn2.microsoft.com/en-us/library/ms996942(d=printer).aspx
6. OASIS Identity Federation (2007) WS-reliablemessaging specification. WS-SecureConversation specification, OASIS
7. OASIS Identity Federation (2011a) Liberty alliance project. Available at http://projectliberty.org/resources/specifications.php
8. OASIS Identity Federation (2011b) Profiles for the OASIS security assertion markup language (SAML) V2.0. Available at http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=security. Accessed 19 Feb 2011
9. Anonymous (2010) Standard for naming active entities on DoD IT networks, version 3.5

10. Simpson WR, Chandersekaran C, Trice A (2008) The 1st international multi-conference on engineering and technology innovation, cross-domain solutions in an era of information sharing, vol I. Orlando, FL, pp 313–318
11. Chandersekaran C, Simpson WR (2008) World Wide Web consortium (W3C) workshop on security models for device APIs, the case for bi-lateral end-to-end strong authentication, London, England, p 4
12. Simpson WR, Chandersekaran C (2009) In: 2nd International multi-conference on engineering and technological innovation, vol 1, information sharing and federation, Orlando, FL, pp 300–305
13. Chandersekaran C, Simpson WR (2011) In: 16th international command and control research and technology symposium: CCT2011, an agent based monitoring system for web services, vol II. Orlando, FL, pp 84–89
14. Simpson WR, Chandersekaran C (2011) In: 1st international conference on design, user experience, and usability, part of the 14th international conference on human-computer interaction (HCII 2011), a multi-tiered approach to enterprise support services, Orlando, FL, p 10
15. Simpson WR, Chandersekaran C, Wagner R (2011) Lecture notes in engineering and computer science. In: Proceedings world congress on engineering and computer science 2011, vol I, high assurance challenges for cloud computing, San Francisco, pp 61–66

Chapter 24
The ACROSS Integrity Model

Armin Wasicek

Abstract In this chapter we discuss the application of integrity models in a mixed-criticality system to enable the secure sharing of information. The sharing of resources and information in computer systems enables cost savings. The major technical challenge of these systems is simple: low criticality applications must be prevented from interfering with high criticality ones which execute in the same system. An example for such an integrated architecture is the the ACROSS MPSoC architecture which facilitates the implementation of hard real-time systems. We present an integrity model for the secure exchange of information between different levels of criticality within ACROSS. Our approach is based on Totel's integrity model which proposes to upgrade information from low to high by rigorously validating this information. We were able to show that the encapsulation mechanisms of the ACROSS architecture support the implementation of the proposed integrity model.

Keywords ACROSS architecture · Integrity model · Mixed-criticality · Resource sharing · Security policy · Totel's model

24.1 Introduction

Mixed-criticality systems integrate applications with different levels of safety and security in a single computer system. The challenge in mixed-criticality systems is to prevent faults and intrusions that propagate from applications with lower criticality levels to applications having a higher criticality level. Therefore, low criticality applications are usually prohibited from communicating to ones having

A. Wasicek (✉)
Vienna University of Technology, Institute for Computer Engineering,
Treitlstrasse 3/3 1040 Vienna, Austria
e-mail: armin.wasicek@tuwien.ac.at

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_24,

a higher criticality. The rules defining these communication flows are called integrity models and they guarantee a proper way of communication in the safety and security domains.

One solution is to build all software at the highest criticality level. This does not only increase complexity, but system development also becomes very expensive. For example, a high criticality application might want to read a sensor. Using sensors with the same (high) criticality level will most likely cause a higher cost than using sensors with a lower criticality level. Methods have to be researched to facilitate the safe and secure sharing of information between criticality levels and to enable the simple and cost-efficient implementation of a mixed-criticality system.

In real-time systems, upgrading data from lower level to be usable in higher levels requires *maintaining consistency in time and space* between different data sets.

The research on the theoretic background of integrity models reaches back to the seventies. Models like Bell-LaPadula and Biba are taught in undergraduate security courses. These models define rules how information may be exchanged between criticality levels. However, in many cases these rules are too restrictive. A more recent model is Totel's model, which provides a formal foundation to enable a more flexible information sharing.

Our approach encompasses the specification and the implementation of integrity model based on Totel's model. We explored the design space for such an model in [21]. We explicitly address real-time constraints by building on the ACROSS architecture which implements a time-triggered System-on-a-Chip (SoC) for application in real-time systems. Our main contributions center around the design of a *Validation Middleware* to check and upgrade information:

- Implementation of a secure information sharing system within the context of the ACROSS MPSoC architecture
- Analysis of different anomaly detection algorithms employed to recognize modifications and voting strategies to upgrade information
- Evaluation of the system in a vehicular simulator

Levels are usually defined according to standardized classifications like Safety Integrity Levels [8] or FIPS 140–1 [10] for security. In this work, we *focus mainly on the security aspects of a mixed-criticality system.*

24.2 Basic Concepts and Related Work

In this section we discuss the fundamental concepts of replica determinism, voting, the ACROSS Multi-Processor System-on-a-Chip (MPSoC) architecture, and related integrity models.

24.2.1 Replica Determinism and Voting

In a *deterministic* computer system, it is possible to predict a future state given its initial state and all timed future inputs. Determinism is a binary system property, i.e., it is either present or not in any computer system. When designing real-time systems, determinism is a particularly desired property, because it

- *simplifies the cognitive effort* required to understand the behavior of the component,
- facilitates the *masking of faulty channels*, and
- *increases testability*, because repeating the same test will produce the same results.

Determinism is lost in a computer system by failures or by non-deterministic design constructs like depending on randomness for decisions, using non-deterministic programming languages, preemptive scheduling, or inconsistent message order in replicated channels [12].

We call a system *replica deterministic*, if *it can be guaranteed by design that all correctly operating replicated components will visit the same states at about the same time, i.e., they will take the same decisions at all major decision points* [11]. This definition requires that the nodes are time-synchronized in order to make their internal states comparable. We observe the internal state of a component through the output messages produced. This is also called a component's behavior. The intended behavior is the service a component delivers [1]. Therefore, two replica-deterministic components (i.e., replicas) will produce the same service within some defined temporal bounds. A difference in the behavior of two replicas can be used an input for the fault detection.

When designing and implementing an *active redundancy* scheme, the presence or absence of replica determinism has a profound impact on the type of voting strategy that can be selected. In an active redundancy scheme, two or more nodes operate concurrently to produce redundant results. Basically, we distinguish between two classes of voting strategies for error detection:

- *Exact Voting* performs a bit-wise comparison of the bit patterns of its inputs. For this techniques replica-determinism is presupposed, because the voting has to be performed over the concurrent outputs of the replicated components. If there is no replica determinism, a temporally ordered message sequence cannot be assumed. As a consequence, race conditions might change the message order and thus affect the voting.
- *Inexact Voting* techniques introduce thresholds to determine approximate equivalence of results. A set of thresholds forms a neighborhood. The goal of an inexact voting technique is to determine, if one input resides within a specified neighborhood of the other value.

Inexact voting techniques are relevant for systems that cannot guarantee replica determinism and systems that observe the physical environment. It is extremely unlikely that two different sensors will collect bit-exactly the same data, even if they are arbitrarily close to one another and sampling the same process. Some analytical method has to be applied to determine, if the data of the sensors is effectively equal, even if not identical.

The difficulty when implementing an inexact voting scheme is the selection of appropriate thresholds to define a neighborhood. *There is no general theory known to help in the selection of these threshold values—so the selection has to be based on heuristics.*

24.2.2 The ACROSS MPSoC Architecture

The ACROSS architecture is the realization of a MPSoC architecture for heterogeneous application cores that provides a deterministic on-chip communication. It targets applications across different domains (e.g., avionics, automotive, industrial). Its particular benefits for application designers encompass fault isolation at the level of application cores, system-wide global time base to coordinate activities in the distributed system, and a temporal firewall interface to decouple computation from communication.

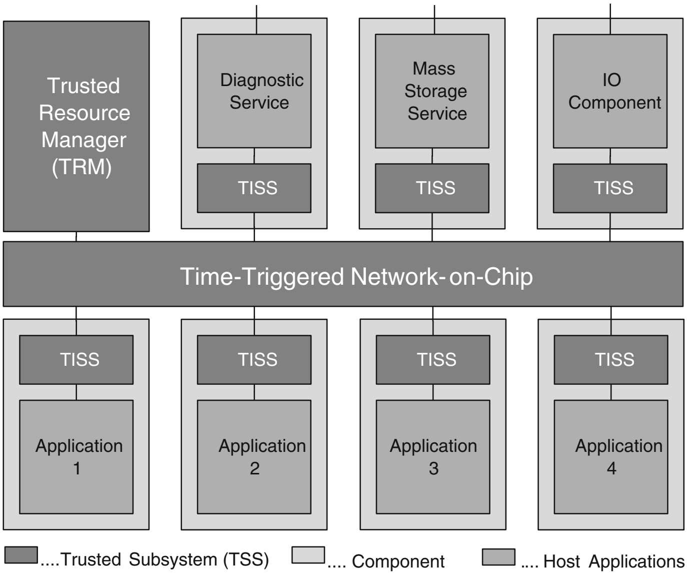


Fig. 24.1 Basic layout of the ACROSS MPSoC

The basic architectural elements of an ACROSS MPSoC [9] are depicted in Fig. 24.1. *Components* are connected through a Time-triggered Network-on-a-Chip (TTNoC). A component is considered to be a self-contained computational element with its own hardware (processor, memory, communication interface, and interface to the physical environment) and software (application programs, operating system), which interacts with other components by exchanging messages. The *TTNoC* transmits messages according to an a priori defined time-triggered schedule. The *Trusted Interface Subsystems* (TISS) are the interaction points between the TTNoC and a component. The endpoints inside the TISS are called *ports*. Every TISS can have multiple input and output ports, but can not send or receive simultaneously. The *Trusted Resource Manager (TRM)* is a dedicated system component that manages the ports and routes on the TTNoC.

A *job* is a constituting element of a Disturbed Application Subsystem (DAS) and forms the basic unit of computation. It interacts with other jobs through the exchange of messages. A DAS is a nearly independent distributed subsystem of a large distributed real-time system that provides a well-specified application service (e.g., a power-train system or a multimedia system). Jobs are allocated to components.

An *encapsulated communication channel* is an unidirectional data channel which transports messages at pre-defined points in time. The encapsulation mechanisms prevent temporal and spatial interference between different components. For instance, delaying or overwriting a message is not possible.

In an ACROSS MPSoC, encapsulated communication channels are established between TISSs through the time-triggered message transfer of the TTNoC. TTNoC accesses are arbitrated between different components through a Time-Division Multiple Access (TDMA) scheme. Every component implements its own local memory. Because no component can neither directly interfere with communication (e.g., by deliberately sending messages at certain instants), nor change a channel's configuration (these are exclusively managed by the TRM), an ACROSS MPSoC implements *segregation in the temporal and spatial domain.*

The encapsulation mechanism of the ACROSS MPSoC are not only beneficial to enforce its dependability properties, but also to efficiently implement security mechanisms. The chapter in [22] discusses the capabilities of the ACROSS architecture to fulfill the basic requirements of a Multiple Independent Levels of Security (MILS) system [6]. The Trusted Subsystem (TSS) of the ACROSS architecture encompasses TTNoC, TISS, TRM and implements a *Separation Kernel* that isolates processes in separate partitions (components).

Following this reasoning, the chapter comes to the conclusion that the ACROSS MPSoC implements a *Trusted Computing Base (TCB)* which is a *small amount of software and hardware that security depends on and that we distinguish from a much larger amount that can misbehave without affecting security* [14].

24.2.3 Totel's Integrity Model

Integrity models have been used for a long time in computer systems. Historically relevant integrity models are the Bell-LaPadula model for confidentiality [4] and the Biba model for integrity [5]. It is common to most integrity models to vertically subdivide a system into integrity levels I that are related with a partial order relation ($\leq$). The system designer assigns the tasks, applications, or subsystems (i.e., objects O) to a particular integrity level. Formally, the function $il : O \rightarrow I$ associates an integrity level to the system's components. The integrity model then defines how the components (e.g., subjects and objects) may interact. For instance, the Biba model defines that a subject should not be allowed to read an object of a higher integrity level (no write up), and to not read one of a lower level (no read down).

Totel's model [20] is a more recent development related to the Biba model. Contrary to Biba, where subjects access objects, Totel's model has only one kind of entities called objects. These objects provide *services* that can be requested by a client. Each object is classified within a particular integrity level that indicates to what degree it can be trusted and what its dependability requirements are. If an object creates a message, this message inherits its creator's integrity level. On the reception of a message, rules are applied to check if this message is valid or not.

The concept of a Validation Object (VO) is central to Totel's model. A VO *takes low level inputs and runs fault tolerance mechanisms to produce high integrity outputs* [20]. It reflects the circumstance that information flows that would be normally regarded as illegal, are needed for a flexible application design. The solution that Totel's integrity model advocates is to increase the trustability of the information contained in lower level objects to make the information usable at higher levels without corrupting higher level objects. The fault-tolerance mechanisms to implement this upgrading are strongly application-dependent. In this chapter, we present a case study to highlight one promising application of VOs in an automotive context.

24.3 Secure Information Sharing Subsystem

In this section we describe how to enable the secure sharing of information between different levels of criticality within an ACROSS MPSoC.

24.3.1 Definitions and Notations

We define our integrity model as a tuple $\langle J, C, M, I, il \rangle$:

- a set J of jobs

- a set C of components of an MPSoC
- a set M of messages on the TTNoC
- an ordered set I of integrity levels, with a partial order relation ($\leq$) between its elements
- $il : C \rightarrow I$, an association of a component to an integrity level

In addition, we will use these two functions:

- $val : M \rightarrow M \cup \{\varepsilon\}$, a validation function
- $cr : M \rightarrow C$, an association between messages and their creating components

24.3.2 The ACROSS Integrity Model

The instantiation of an ACROSS MPSoC provides a basic platform to implement a multi-level secure system with different levels of criticality. The TSS of the ACROSS MPSoC acts as TCB to encapsulate communication flows and to segregate criticality levels and components. The encapsulation within the architecture is achieved by strictly assigning one job to one component. Integrity levels are per definition assigned to components.

Rule 1 *One job is assigned to exactly one component:*

$\exists! j \in J, \exists! c \in C,\ j \rightarrow c$

The Unidirectionality property of an encapsulated communication channel is essential to implement Biba's rules:

Rule 2 *Information flow is allowed only between components on the same or to a lower criticality level:*

$\forall (c_1, c_2) \in C \times C,\quad c_1\ send\ c_2 \Rightarrow il(c_1) \geq il(c_2)$

Figure 24.2 depicts an exemplary instantiation of the ACROSS integrity model. Each component on the MPSoC is assigned a desired integrity level. Next, each job is allocated to a component (rule 1). The arrows represent the encapsulated communication channels connecting the components (using rule 2). They all pass through the TSS. If a communication path is required from a lower level component to a component at a higher level, a Validation Middleware (VaM) has to be placed within the receiving component. The VaM upgrades the information from several sources by applying a validation function.

Rule 3 *Information flows from lower to higher integrity levels must pass through a VaM:*

$C_{low} send c_{high} \Rightarrow$

1. $il(c_i) < il(c_{high}), c_i \in C_{low} \subseteq C$
2. $M_{in} = \{m_i \in em(C_{low})\}$
3. $M_s \subseteq M_{in} | m_i, m_j \in M_s, |z(m_i) - z(m_j)| < \delta, m_i \neq m_j$
4. $val(M_s) = M_v, \forall m_i \in M_v, m_i \neq \varepsilon$

Note that criterion 1 in rule 3 states that all incoming messages are received within a guaranteed upper bound, i.e., they are ordered within a specified time interval. Criteria 2 and 3 postulate that each message origins by a different, diverse or redundant sending component. Finally, criterion 4 requires that validation functions are only applied to transmissions from a lower level to a higher one.

Basically, the definition of the ACROSS integrity model relies on the concepts of Totel's model. Because Totel's model is thought to be used in a single processor system, it has to be adapted for the use within an MPSoC architecture like ACROSS.

No Integrity Kernel Required: The original model requires upward communication to pass through an integrity kernel. In ACROSS, all communication is passing through the TSS which is considered to be a TCB. Moreover, because all information flows are clearly defined and rigorously checked at run-time by the TRM, there is no possibility that a covert channel can exist.

No Multi Level Objects (MLO) Defined: In the ACROSS integrity model, the concept of an MLO does not fit into the system design. The encapsulation of architecture enforces a strict assignment of one object to one component and one integrity level. The architecture by itself cannot guarantee that concurrent invocations of an MLO within a single component do not influence each other, because a component is considered as an atomic unit. Accesses to the TISS ports from within a component are not arbitrated by architectural means. However, if additional segregation mechanisms are implemented within a component (e.g., through a partitioning operating system), concurrent access to the TISS can be mediated and MLO can be supported.Therefore, the distinction from Single Level Object (SLO) and MLO is not necessary.

No Integrity Checks Required: The invocation model of Totel assumes that object invocations are implemented through message-passing. A message is on the same level of integrity as its creator and carries a label indicating its level. In ACROSS these labels are not required, because messages can only be transmitted via encapsulated communication channels. The rules of the ACROSS require that channels are only established between objects on the same level. These rules have to be applied at design time and the strict adherence to these rules is enforced by the TRM.

No Read–Write Rule: Each port in the TISS can be used either for reading or for writing, but it is not possible to perform both actions at the same time. Therefore, the read–write rule from Totel's model can be dropped in ACROSS.

24.4 Validation Middleware

This section explains how the previously proposed integrity model is used in combination with the ACROSS architecture. We describe the specification and implementation of the validation function in form of a middleware layer. This VaM is then used to upgrade the criticality level of upstream information flows.

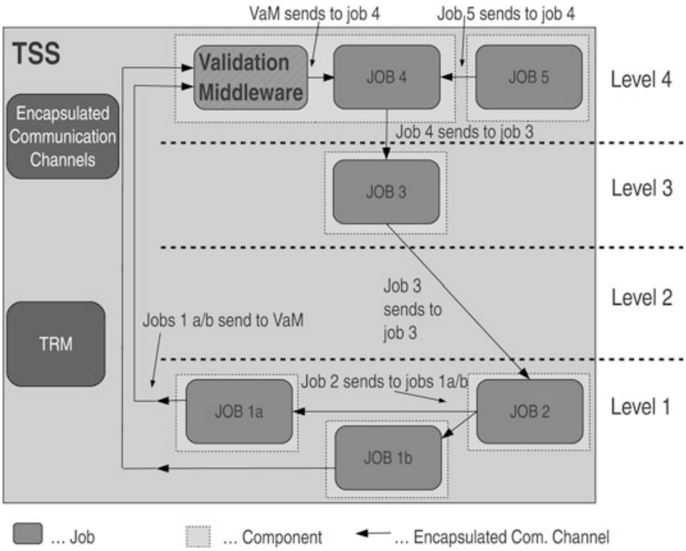


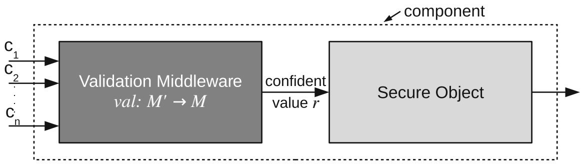
Fig. 24.2 The integrity model for ACROSS

Each component is assigned a level of criticality. Jobs are allocated to components and they produce messages that are then transmitted over encapsulated communication channels via the TSS. In case an upgrade of information between criticality levels is required (shown in Fig. 24.2 between job1a/b and job 4), a VaM is inserted. The VaM gathers and processes information from redundant and diverse sources to produce the upgraded information.

Figure 24.3 depicts a VaM's basic information flow which provides *N* different and potentially diverse inputs, and a single output *r*. The primary purpose of a VaM is to evaluate the validation function *val*. In order to produce a meaningful output the input channels need to be completely independent from each other. An approach to achieve this independence is called N-Version Programming (NVP) [2].

The main goal of NVP is to eliminate faults which are introduced during design and implementation of a system. The NVP approach uses functionally equivalent programs that are independently generated from the same initial specifications. This independence of each version is also called design diversity. The main purpose of such required diversity is to eliminate the commonalities between different programming efforts, because they have the potential to cause related faults. Diversity is realized by using different teams, algorithms, programming languages, tools and

Fig. 24.3 Validation middleware



computational platforms. The output of each version needs to be compared by a output selection algorithm or a voting algorithm. This is a similar problem to that of realizing a VaM.

The most common and intuitive way of comparing the multiple inputs of a VaM is a majority voting algorithm, but this creates some difficulties [15]. In some situations, the inputs need to be compared and determined to be correct, even if the input data differs. This raises the need for an inexact voter. Hence, inexact voting algorithms represent a solution for the implementation of a VaM. The requirements for these algorithms follow below.

24.4.1 Proposed Anomaly Detection Algorithms

The algorithms used to implement a VaM should be able to detect errors and prevent wrong values from propagating into the secure area of the application. The most important requirement of a VaM is that it has to be certified at the same integrity level as the application in the component itself. Therefore, it is desirable to keep this middleware simple and reusable. All algorithms should run online, possibly without knowing anything from the previous run. This kind of memoryless algorithms alleviate verification because the number of possible outcomes of a set of input data is always the same. The sources of data (i.e., sensors) often induce noise or even omit a value. For this purpose, they should provide filter mechanisms by pre-processing the input data. The prevalent resource constraints in embedded systems call for a light-weight approach with a low memory and computation footprint.

We propose a set of four anomaly detection algorithms that comply with these requirements. Table 24.1 summarizes their time and space complexity properties. We present an analysis of these algorithms in an automotive environment in Sect. 24.6.

*k*th *Nearest Neighbor with Δ* $-$Value: This algorithm calculates an anomaly score of each data instance, by counting the number of nearest neighbors (k) that are not more than a distance d apart from the given data instance [19]. After assigning an anomaly score to every data instance, it needs to be defined for which score the data is anomalous. Hence, defining the distance d is an application-dependent challenge. Mostly, this calibration is done by using test data. In this chapter, the algorithm is used with a small data set and therefore the majority of all

Table 24.1 Validation middleware algorithm overview

	Time compl.	Space compl.
*k*th near. neighbor w. Δ–Value	$O(n^2)$	$O(n)$
Probabilistic boxplot method	$O(n\ log(n))$	$O(n)$
Histogram method	$O(n)$	$O(n)$
Single-linkage clustering	$O(n^2)$	$O(n^2)$

values is used. If the anomaly score is higher or equal to the score of the majority of data instances, normal behavior is attested.

Probabilistic Boxplot Method: This algorithm produces a boxplot diagram by using the input data samples [3]. This kind of diagram is normally used for graphical illustration of statistical data but it can be used as a method to detect anomalies. A boxplot is fully defined by five values. The mean, median, 25th percentile, 75th percentile and the interquartile range. A sorted input data set of values is used to derive these parameters. The most significant factor for detecting anomalies in a data set is the interquartile range ($IQR = Q_{.75} - Q_{.25}$). If a data sample is outside the range of $x_m \pm (IQR \times 1.5)$ this sample is marked as an anomaly. The factor of 1.5 is not a fixed quantity, but tunable.

Histogram Method: The histogram based anomaly detection algorithm is one of the simplest non-parametric statistical technique used in this area. A histogram is a graphical illustration of a frequency distribution. It is based on the classification of data into bins of fixed or variable width. The size of the bins represents the relative frequency of the data inside the box. The algorithm basically consists of two steps. The first step involves building a histogram based on the input data. In the second step, each data sample gets assigned to a bin of the histogram. After this is done, the bin with a majority of the values is selected and returned.

Single-Linkage Clustering: Clustering tries to find a structure in a collection of unlabeled data. A cluster is a set of objects, which are similar to each other. Single-Linkage clustering declares data instances belonging to clusters whose size is below a certain threshold value as being anomalous. It belongs to the class of hierarchical clustering methods. Generally, it works bottom–up. Each data instance starts in its own cluster and pairs of clusters are merged and moved up the hierarchy. To be able to decide which cluster should be merged, the distances between the data instances are calculated [16].

24.5 Automotive Case Study

In this section we present an automotive case study as a proof-of-concept of the presented integrity model and to evaluate different algorithms implementing the case study.

We choose two related car functions with mixed-criticality requirements for safety and security for our case study:

Odometer Subsystem: An *odometer* computes and stores the current mileage counter of a vehicle. In order to log the distance covered by the car, the current speed needs to be sampled and multiplied with the time elapsed since the last measurement. The accumulated results of this computation is the total distance traveled. This subsystem has a high security requirement, because a vehicle's resale value depends largely on this value. Odometer fraud is the illegal practice of rolling back odometers to make it appear as if vehicles have lower mileage than they actually have [17].

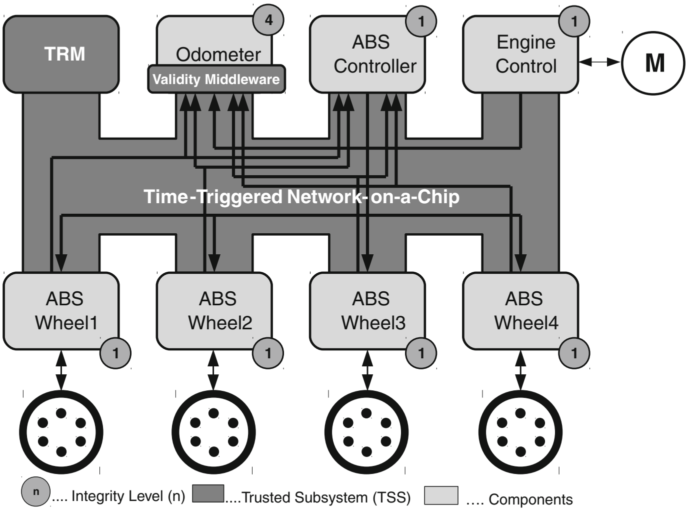


Fig. 24.4 Allocation of automotive subsystems to ACROSS MPSoCs

ABS Subsystem: The second subsystem is an Anti-Blocking System (ABS). The ABS prevents wheel lock-up during heavy braking [7]. The ABS controller detects the wheel lock-up as a sharp increase in wheel deceleration. In our case study, a four channel, four sensor system is used, which has a speed sensor on each wheel and separate valves to apply a brake force to each wheel. Security requirements are not common in ABSs.

Integration of Subsystems: The traditional setup of a car is to deploy the ABS and odometer subsystems in a federated approach by physically disjoint systems. The advantage of such an approach is that the segregation between subsystems is high. However, this comes at a high cost of many redundant units. Integrated architectures like ACROSS aim to deliver similar segregation properties present in a federated system and simultaneously reduce size, weight, power consumption of the hardware while reducing complexity and increasing re-usability of the software [18]. Figure 24.4 depicts the mapping of the automotive subsystems onto the chip.

To simulate attacks on the anticipates system, we developed a hardware-in-the-loop system based on the Open Racing Car Simulator (TORCS)[1] to evaluate our concepts. TORCS provides a realistic physics environment, car models, race tracks, and an interface to program robots, which are programs that steer the vehicle.

We use the attack tree method to assess potential attacks on the odometer task (see Fig. 24.5). There are three main parts of the odometer subsystem that are viable attack targets:

[1] http://torcs.sourceforge.net/

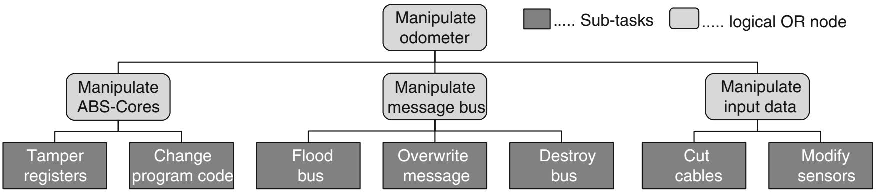


Fig. 24.5 Attack model for the odometer subsystem

1. *Manipulate jobs:* This can be achieved either by overwriting the registers that hold the speed values of the sensors, or by manipulating the program running on the ABS components of the wheels (e.g., downloading a modified version). This branch of the attack tree can be prevented by software security mechanisms which are out of scope of this chapter.
2. *Manipulate communication system:* This can be carried out through a flooding attack to block the communication or to periodically overwrite messages to induce wrong speed information. These attacks are prevented by the encapsulation mechanisms of the architecture.
3. *Manipulate input data:* Five redundant and diverse sensors are used to provide the current speed of the vehicle. The simplest manipulation is to cut the cables which connect the sensors with the communication system (i.e., create a stuck-at-value). Another way is to tamper with the measurement of the installed sensor. If the odometer depends on a single sensor (as it is the state-of-the-art), it is attackable. If the odometer depends on five redundant and diverse sensors like in our case study, its manipulation is hindered. If in addition, these sensors are used in conjunction with the safety-critical part of the system, an attacker has might be refrained from tampering the sensor devices of the ABS subsystem.

The simulation environment supports software-based fault injection by modifying the contents of the shared memory. The following faults can be simulated:

- The wheel sensor can be disabled, which simply freezes the value at the last measured value.
- The current speed of a sensor can be set to a fixed level.
- The ABS subsystem can be enabled or disabled.
- A fault injection schedule containing the name of the faulty sensor, the start- and end times and the modified speed value can be defined. This schedule is executed during run-time.

24.6 Evaluation

We use the simulation environment presented in the previous section to evaluate the proposed secure information sharing system. We assume the rightmost branch (‘Modify Sensors’) of the attack tree, hence, an attacker is tampering with the

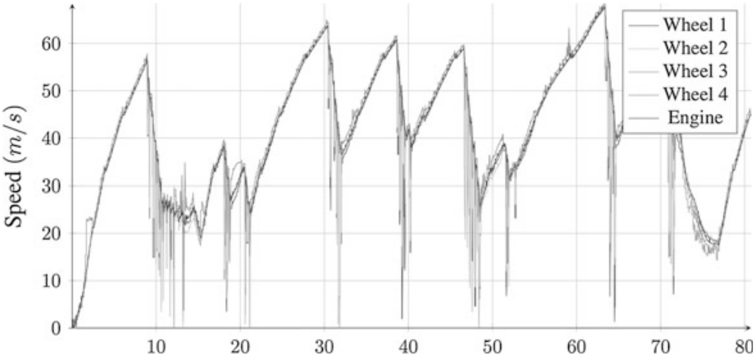


Fig. 24.6 All five speed values during one lap of a race

Table 24.2 Percentage of results determined as valid

	No faults injected (%)	Faults injected (%)
*k*th Nearest Neighbor	93.58	88.49
Boxplot	99.96	99.98
Histogram	96.04	89.19
Singe-Linkage Clustering	100.00	100.00

physical interface of a sensor. Fig. 24.6 depicts the evolution of a vehicle's speed during a sample race.

The deviation of the speed sensor values is the main testing criteria for the anomaly detection algorithms used in the VaM. The algorithms need to find a majority of speed values with a small variance and eliminate possibly wrong or inaccurate values. The optimal behavior of an algorithm is that it always finds a manipulated speed value, which would increase or decrease the current value of a sensor and therefore change the current mileage. Secondly it should neglect values from wheels which are currently locked up or spinning, because these values would also influence the computed mileage.

We conducted several experiments (with and without fault injection) to study the proposed algorithms. Table 24.2 gives a brief summary of the outcome of the experiment runs with and without fault injection.

24.7 Conclusion

In a mixed-criticality system with multiple security levels the induction of faults in a secure component is detected and prevented through the deployment of a VaM. The VaM is used to upgrade the integrity of this information flow. Therefore, the VaM needs diverse and redundant inputs to upgrade the information. We propose to use inexact voting based on anomaly detection algorithms which do not aggregate values between subsequent executions. This is a particular requirement to facilitate certification. However, in extreme driving scenarios (e.g., braking), when all sensors show different readings, these algorithms do not work accurately.

Most important, to realize the proposed secure information sharing system, is the availability of diverse and redundant inputs that are then used to upgrade information. A particular strong point of our design is that we are able to separate the execution platform configuration (including the configuration of the integrity model) and the computation: platform configuration can be certified independently from jobs. Summarizing, our integrity model to securely exchange information between different criticality levels can be efficiently implemented in an integrated architecture like ACROSS. Certifiable anomaly detection algorithms which actually perform the upgrade require further research.

Acknowledgments This document is based on the ACROSS project in the framework of the ARTEMIS program. The work has been funded in part by the ARTEMIS Joint Undertaking and National Funding Agencies under the funding ID ARTEMIS-2009-1-100208.

References

1. Avizienis A, Laprie JC, Randell B, Landwehr C (2004) Basic concepts and taxonomy of dependable and secure computing. IEEE Trans Dependable Secure Comput 1(1):11–33
2. Avizienis AA (1995) The methodology of N-version programming. In: Lyu M (ed) Software fault tolerance. Wiley, New York, pp 23–46
3. Banerjee A, Kumar V (2009) Anomaly detection: a survey. Technical report, ACM computing survey
4. Bell DE, LaPadula LJ (1975) Computer security model: unified exposition and multics interpretation. Technical report, MITRE Corp., Bedford
5. Biba KJ (1977) Integrity considerations for secure computer systems. Mitre Corporation, technical report
6. Boettcher C, DeLong R, Rushby J, Sifre W (2008) The MILS component integration approach to secure information sharing. In: Proceedings of the 27th digital avionics systems conference (DASC). IEEE/AIAA
7. Burton D, Delaney A, Newstead S, Logan D, Fields B (2004) Effectiveness of ABS and vehicle stability control systems. Technical report, Royal Automobile Club of Victoria (RACV) Ltd
8. Commission I.E.: IEC 61508 (2005) Functional safety of electrical/electronic/programmable electronic safety-related systems. In: 1st IEEE automotive electronics conference, pp 7–13
9. El-Salloum C, Elshuber M, Höftberger O, Isakovic H, Wasicek A (2012) The ACROSS MPSoC - a new generation of multi-core processors designed for safety-critical embedded systems. In: Proceedings of the 15th euromicro conference on digital systems design (DSD)
10. Evans DL, Bond PJ, Bement AL (2001) Security requireents for cryptographic modules. Federal Information Processing Stabdards Publication (Supercedes FIPS PUB 140–1)
11. Kopetz H (1995) Why time-triggered architectures will succeed in large hard real-time systems. In: FTDCS, pp 2–9
12. Kopetz H (2011) Real-time systems: design principles for distributed embedded applications, 2nd edn. Springer, Berlin

13. Laarouchi Y, Deswarte Y, Powell D, Arlat J (2003) Connecting commercial computers to avionics systems. In: 28th digital avionics systems conference pp 6.D.1-(1–9)
14. Lampson B, Abadi M, Burrows M, Wobber E (1992) Authentication in distributed systems: theory and practice. ACM Trans Comput Syst 10(4):265–310
15. Lorczak PR, Caglayan AK, Eckhardt DE (1989) A theoretical investigation of generalized voters for redundant systems. In: Digest of papers FTCS-19: the nineteenth international symposium on fault-tolerant, computing, pp 444–450
16. Matteucci M (2000) Hierarchical clustering algorithms. Available at: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchic% al.html
17. Morgan C (2002) Preliminary report: the incidence rate of odometer fraud. Technical report DOT HS 809 441, national highway traffic safety administration (NHTSA)
18. Obermaisser R, El Salloum C, Huber B, Kopetz H (2009) From a federated to an integrated automotive architecture. IEEE Trans Comput Aided Des Integr Circ Syst 28(7):956–965
19. Subramaniam S, Palpanas T, Papadopoulos D, Kalogeraki V, Gunopulos D (2006) Online outlier detection in sensor data using non-parametric models. In: Proceedings of the 32nd international conference on very large data bases (VLDB), pp 187–197
20. Totel E, Blanquart JP, Deswarte Y, Powell D (2000) Supporting multiple levels of criticality. ESPRIT project 20716: GUARDS
21. Wasicek A, Mair T (2012) Secure information sharing in mixed-criticality systems. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, IAENG, pp 23–29
22. Wasicek A, Salloum CE (2010) A system-on-a-chip platform for mixed-criticality applications. In: Proceedings of 13th IEEE international symposium on object/component/service-oriented real-time distributed computing (ISORC)

Chapter 25
Modeling of the Stress Distribution in Temporomandibular Joint with Subtotal Replacement

Josef Daněk, Tatjana Dostálová, Milan Hubáček, Nima Mahdian and Jiří Nedoma

Abstract The temporomandibular joint (TMJ) is one of the most complicated joints of the human skeleton. It is a complex, sensitive and highly mobile joint which works bilaterally so each side influences the contralateral joint and because of this the distribution of the stresses is changed in the healthy joint as well. Detailed knowledge about function these are necessary for clinical application of temporomandibular joint prosthesis and also help us estimate the lifetime of the prosthesis a possibilities of alteration in the contra lateral joint components. The mathematical model of TMJ with replacement is based on the theory of semi-coercive unilateral contact problems in linear elasticity and on finite element approximation. The geometrical model of the TMJP was created using the dataset of axial computer tomography. The main objective of our investigation is to discuss numerical results for model of TMJ with subtotal prosthesis and to characterize processes in joint as well as in replacement.

J. Daněk (✉)
European Centre of Excellence NTIS - New Technologies for the Information Society,
University of West Bohemia, Univerzitní 8, 306 14 Plzeň, Czech Republic
e-mail: danek@kma.zcu.cz

T. Dostálová · M. Hubáček · N. Mahdian
Department of Paediatric Stomatology, Charles University, University Hospital in Motol,
V Úvalu 84, 15006 Prague, Czech Republic
e-mail: Tatjana.Dostalova@fnmotol.cz

M. Hubáček
e-mail: Milan.Hubacek@fnmotol.cz

N. Mahdian
e-mail: mahdian@centrum.cz

J. Nedoma
Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod
Vodárenskou věží 271/2, 18207 Prague, Czech Republic
e-mail: nedoma@cs.cas.cz

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_25,

Keywords Contact problem · Cyst · Finite element method · Linear elasticity · Mathematical modeling · Maxillofacial surgery · Subtotal prosthesis · Temporomandibular joint

25.1 Introduction

The temporomandibular joints function symmetrically and this harmony allows biting, chewing and speaking. There are two types of movement: (1) rotary movement, (2) sliding movement. Movements are mostly combined together resulting in the following jaw movements. Because the TMJ is a bilateral joint, function or change of one side influences the contralateral side. In our case the right TMJ is resected and replaced by subtotal prosthesis. During the surgery the medial pterygoid muscle, masseter muscle and temporal muscle were cut off and resutured to the replacement (subtotal prosthesis UniLOC Recon Plate 2.4-Synthes, Inc., West Chester, USA). Temporomandibular disorders (TMDs) is a generic term and may occur for many reasons involve among others pathological processes in condyle of mandible. Temporomandibular disorders (TMDs) are a term embracing a number of clinical problems that involve the masticatory musculature, the temporomandibular joint and associated structures, or both. These disorders are accompanied by pain in the masticatory muscles, in the TMJ, and in the associated hard and soft tissues. Other symptoms include limitation or deviation in the mandibular range of motion, TMJ sounds, and/or headaches and facial pain.

In this case we deal with patient, who underwent surgery because of a large cyst (see Fig. 25.1a) on the right mandibular ramus from the condyle to the angle of the mandible and reconstruction by partial join prosthesis (see Fig. 25.1b) and as further discussed in the case report, now two years after surgery is the patient with

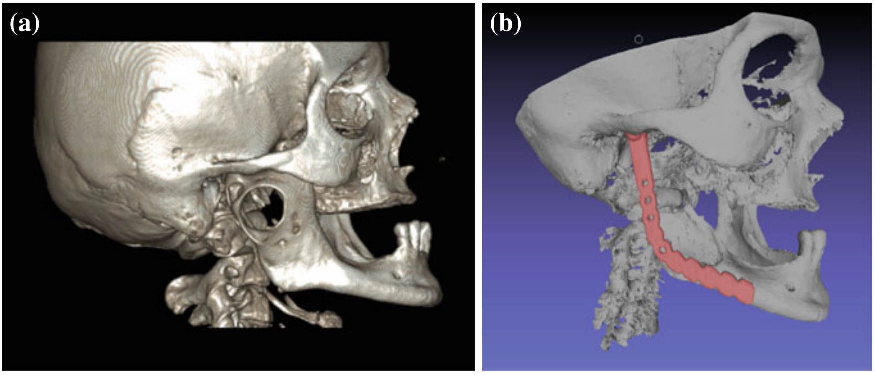


Fig. 25.1 Comparison of the jaw: **a** before and **b** after the surgery

minimal discomfort. Applying the subtotal TMJ replacement biomechanics of the joint system changed. An imbalance will result in a failure of the function and integrity of the TMJ. Therefore, the aim of the mathematical model of TMJ and TMJ prosthesis (TMJP) functions is to establish conditions for preventing any imbalance of the harmony and potential destruction of the TMJ and TMJP. TMJ is strained by pressure and traction, the contact surfaces of TMJ lead to separate in the case of traction and to press in the case of pressure action. Therefore, it is important to mathematically simulate and to analyze the different behavior of each joint during jaw movements, and above all, during nonsymmetrical movement after the surgery. Since the mathematical model allow us to evaluate the application of mechanical and biomechanical aspects of TMJ on prosthesis of TMJ (TMJP). The construction of TMJP and its application by surgical treatment must satisfy or be as much as possible close to human physiological biomechanical parameters, only then the TMJP for our patient will function for a long time without great difficulties. This is the aim of our study for the discussed patient with the large cyst of mandible ramus. Since the patients glenoid fossa was in a good condition, the reconstruction of the right TMJ was made by using the subtotal replacement only. Therefore, the object of our study was a patient after implantation of a subtotal TMJ replacement after resection of right mandible ramus due an extensive cyst. We focused on evaluation of the present and future function of her reconstructed TMJ joint. For this reason we first modeled the healthy 3D model of the mandible, the used data were the data set of axial CT. The results were published in [3, 5]. The initiatory results of stress-strain analysis for model of TMJ with subtotal prosthesis was presented in [2].

25.2 Odontogenic Cysts

Odontogenic cysts are cysts of the jaw which are originated (lined) by an odontogenic epithelium (epithelial tissue). They have an epithelial lining which derive from the epithelial residues of the tooth–forming organ–glands of Serres (rests of the dental lamina), rest of Malassez (rest of the root sheath of Hertwig), or the reduced enamel epithelium.

There are several classifications of the odontogenic cysts, e.g. the World Health Organization (WHO) classification, Shear’s classification or Shafer’s classification. The odontogenic cysts are divided into two groups—the inflamantory and developmental types. Inflamantory types are radicular (or periapical), residual and paradental cysts. Developmental types are odontogenic keratocyst or keratocystic odontogenic tumor. The radicular cyst, the odontogenic keratocyst and the dentigerous cyst are the most important in the stomatological practice, while other types are very rare. For more details see [1, 7, 10].

Since the odontogenic cysts grow within the maxillary bones, they may be causes of bone and tooth resorption, bone fracture, bone expansion or tooth migrations.

The great number of diagnosed odontogenic cysts are the so-called radicular cysts, i.e., OC associated with the roots of non-vital teeth, which are of about 60–75 % of all diagnosed jaw cysts. These cysts arise from the proliferation of the so-called epithelial rests of Malassez. These epithelial rests are the remnants of the root sheath of Hertwig responsible for the formation of the roots of teeth. As it was shown in [11] radicular cysts originate from proliferation of epithelial residues (or epithelial rest). These epithelial cells remain quiescent throughout life, however, cell mediators and signalling molecules released during an inflammatory process may trigger their proliferation. The mechanisms which take place in the pathogenesis of radicular cysts are described in [7, 11], etc. and are based on (1) cyst formations and (2) cyst growth.

25.3 The Model of TMJ with Subtotal Prosthesis

The model for mathematical (numerical) analyses of our patient case was constructed on the basis of real geometry, based on the data from the 3D-CT scan of the destructive cyst on the right ramus mandible, and, therefore, it renders it possible to estimate and to evaluate the future function of both TMJ. Such 3D simulation also brings us new views for evaluation of reconstructive performance in facial skeletal system. To fully understand the response of the glenoid fossa to the prosthesis we need to understand, how the internal forces are distributed through the prosthesis to glenoid fossa and how the changes in right side of joint influence contra lateral joint. Here mathematical modeling of movements of TMJ and distributions of stress-strain fields in operated joint can be used for better understanding of TMJ and its artificial replacement, biomechanical aspects, its function and morphology.

TMJ devices are used as endosseous implants for replacement of each part of temporomandibular joint. For a TMJ implant to be successful is important biocompatibility, low wear and fatigue materials, adaptability to anatomical structures, rigidly stabilized components, corrosion resistant and non-toxic nature. TMJ implant devices can be generally divided to subtotal (partial) or total replacement. The first one is consisted just from one component and it depends, if we want to reconstruct the glenoid fossa (the fossa component) or we want to replace the condyle of mandible (the condyle component). The total replacement consists both the fossa and condylar components. In this area underwent great development custom made total joint prosthesis. It is also because of progress in 3D imaging technology and possibility of mathematic modeling of human skeletal system. Thanks to this is possible by using data from 3D computer model make a

3-dimensional plastic model of the TMJ and associated jaw structures and on this model fabricate a custom-made total joint prosthesis conforming to the patients specific anatomical morphology and jaw interrelationships. Using this technology allows also correction of facial deformities, which are often associated with TMJ disease, in the same operation [13].

In our case we used subtotal replacement of condyle and ramus mandible, because the clinical examination and CT scan didn't show any destruction of glenoid fossa. Also the surgery is not so stressful for the patient and rehabilitation after the surgery is easier and faster. Another benefit of the subtotal replacement is possibility of use before of the end of facial skeletal system growth.

The stress-strain analyses of TMJ and TMJP based on several numerical models and methods, namely the finite element method, were studied by several authors [6, 9].

The mandible, the prothesis and the related parts were approximated by the tetrahedral 4-nodes FE elements. The magnitudes of muscle contraction forces (in N) were estimated with the product of the cross-sectional region of the muscle (in cm^2), the averaged activation ratio, which is taken for all muscles the same of the masseter, and the constant $v = 40$ N/cm^2 [12]. The FEM model simulates the statically loaded mandible in occlusion, where the TMJ is modeled as an ellipsoid-and-socket (or a ball-and-socket) joint. The prosthesis is applied by such a way that the location of the "center" of rotation is steadily fixed. In the compression with the left (healthy) TMJ any muscle forces are neglected.

The frictional force on the contacts between the loads of TMJ (P) right and light joints are approximated by the given frictional forces based on the Coulombian law of friction. Due to the existence of the synovial liquid (fluid) in the TMJ joint, the coefficient of Coulombian friction is very small, so that the frictional forces can be neglected in special cases.

The mathematical model and its numerical solution is based on the theory of semi-coercive unilateral contact problems in linear elasticity [4, 8]. The contact between the condyles and the joint discs are approximated by the unilateral condition.

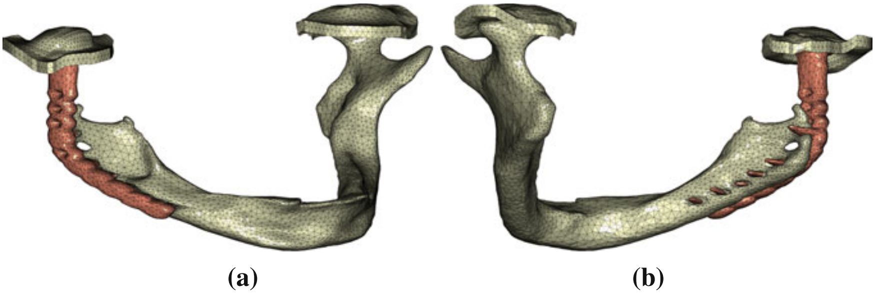


Fig. 25.2 The finite element mesh: **a** front view, **b** back view

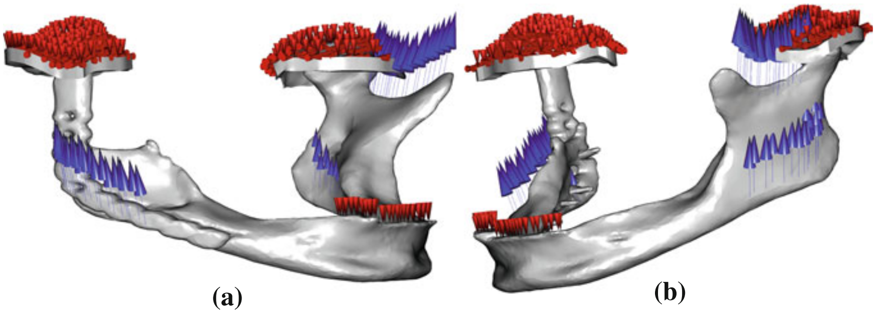


Fig. 25.3 The boundary conditions: **a** front view from right, **b** front view from left

The finite element mesh is characterized by 43,107 tetrahedrons with 12,006 nodes (see Figs. 25.2a, b). The contact boundaries between the condyles of the mandible and the glenoid fossa are approximated by 40 nodes. For the numerical model the following boundary conditions are prescribed: (1) the temporal bone, where the sockets of TMJ are located are fixed; (2) at the upper side of the teeth vertical displacements of about 1mm are prescribed; (3) we have (a) functioning masticatory muscles of the left TMJ acting on the head of the mandible of loads of about $[0.9, -0.6, 2.8] \times 10^6$ Pa for the lateral pterygoid muscle, about $[0.7, -0.6, 2.8] \times 10^6$ Pa for the masseter muscle and about $[0, -0.5, 1.5] \times 10^6$ Pa for medial pterygoid muscle are prescribed; (b) functioning masticatory muscles of the right TMJ acting on the head of the artificial prosthesis of loads of about $[-0.7, -0.2, 2.4] \times 10^6$ Pa for the m. masseter and about $[0, -0.5, 1.5] \times 10^6$ Pa for the m. pterygoideus medialis are prescribed, where the m. pterygoideus lateralis was get out during the surgical treatment (see Figs. 25.3a, b).

For the realization of the numerical solution the COMSOL Multiphysics with the Structural Mechanics Module were used.

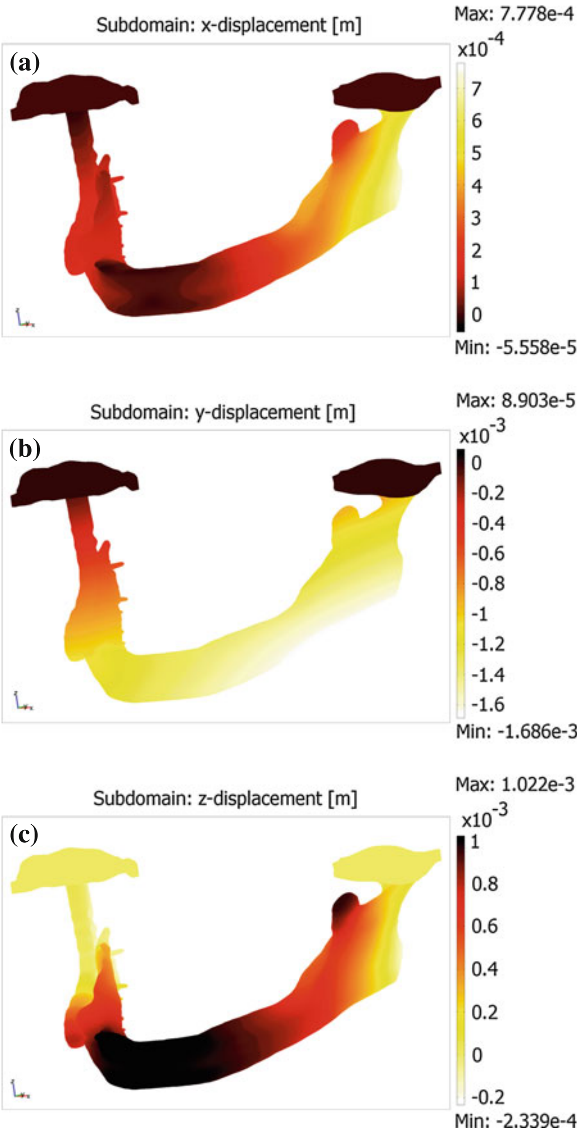
25.4 Results and Discussion

The main objective of this investigation is to introduce a three-dimensional finite element model to calculate the static loading of the TMJP and to characterize processes in the TMJP during its function. A geometrical model of the TMJP was created using the dataset of axial computer tomography (CT).

The values of material parameters are $E = 1.71 \times 10^{10}$ Pa, $\nu = 0.25$ for the bone tissue and $E = 2.08 \times 10^{11}$ Pa, $\nu = 0.3$ for the material of replacement. For the numerical model we set the boundary conditions presented in Figs. 25.3a, b.

At Figs. 25.4a, b the horizontal displacement components in the directions of the x-axis and y-axis tell us about the movements (due to deformation of the

Fig. 25.4 The displacement components in the directions of: **a** the *x*-axis, **b** the *y*-axis, **c** the *z*-axis



mandible) of the loaded mandible in the horizontal plane, moreover, in the consequence of the operated muscles. We see that their effect is greater in the area of the left TMJ joint. At Fig. 25.4c the vertical displacement component is given. It is shown that the minimal value of -2.339×10^{-4} m is in the area of both glenoid fossa, vertical part of replacement and dorsal part of mandible ramus on the left

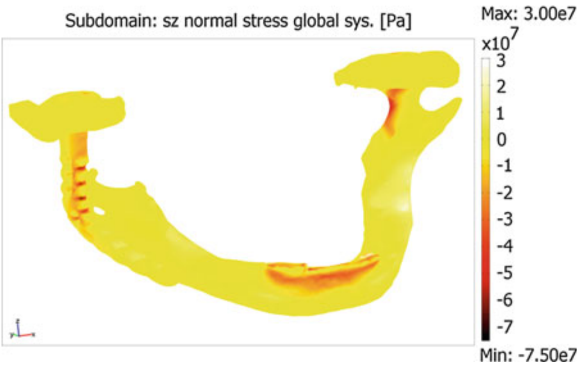
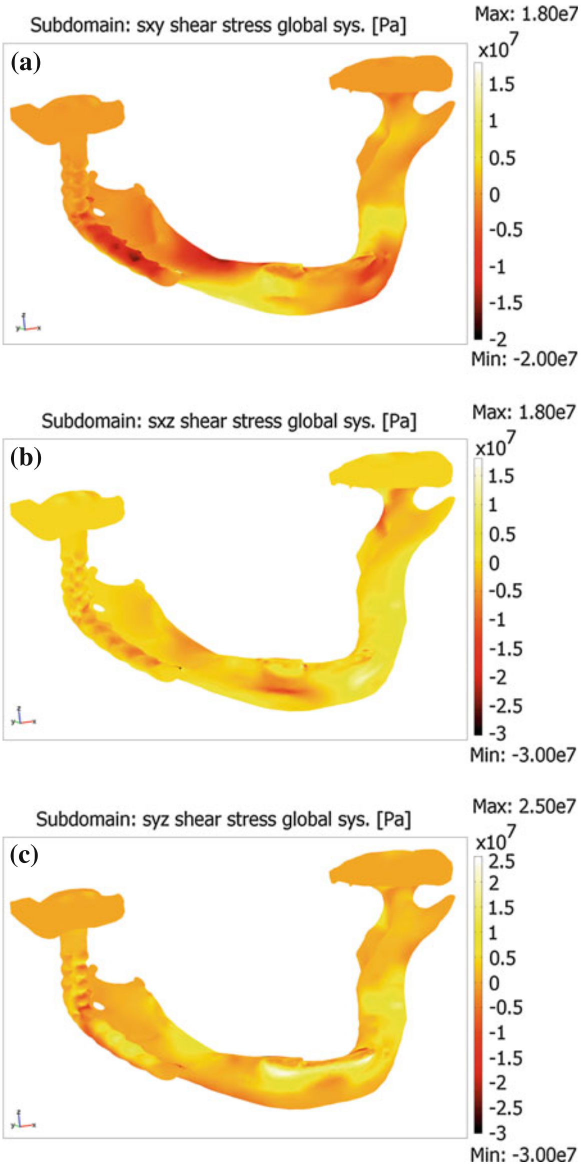


Fig. 25.5 The vertical stress component

(healthy) side and the maximal values are in the area of mandible corpus and coronoid processus of the left side.

At Fig. 25.5 the vertical stress in Pa is presented and at Fig. 25.6a–c the shear stresses in Pa are presented. The von Mises (Fig. 25.7) and the principal stresses (Fig. 25.8) have a great expressive value for specialist in maxillofacial surgery. The von Mises stress is a mathematical combination of all components of both axial and shear stresses. The principal stresses inform us about stresses in directions of the principle axes. They can be usually used to describe the stresses in the studied mandible, and, therefore, they are reasonable indicators, where failures and fractures can later occur. The maximal von Mises stresses in the mandible are located in dorsal part of column mandilble, in the area of the artificial prosthesis and in alveolar processus of corpus mandible. We see that the more informed value have principal stresses. The principal stresses are characterized by pressures, denoted by ↣⨯↢ (red color), and by tensions, denoted by ⟷ (blue color). The mandible in its frontal part is vaulted in the chin elevation that is then passing into the mandible body, where in its frontal part the pressures are observed. In the posterior part (i.e. from the interior = inner side) of the mandible the tensions are observed (see Fig. 25.9a). The upper margin of the mandible body projects (jets) in the alveolar processus with the tooth beds where tensions are observed, while in the lower margin of the mandible, which is rounded off (plumbed), the pressures are indicated (see Fig. 25.9b). The ramus of mandible is closed by the condylar and coronoid processus and that are separated by the mandible incisure. In this area in its anterior part in the healthy TMJ the pressures are observed, while in its posterior part tensions are observed, which is characterized by its bending. In the other side, where is the replacement, the distribution of stresses in the prosthesis is different in comparison with the healthy part of TMJ. The load from the joint is transported as the pressure in the anterior part of the prosthesis, while in its posterior part (side) small amplitude of tensions are indicated. From the distribution of stresses in the mandible, it is evident that the loading of the mandible

Fig. 25.6 The shear stress components: **a** in the plane *xy*, **b** in the plane *xz*, **c** in the plane *yz*



body as well as TMJ is different, but the TMJ is functioned, although distribution of the acting muscles is different, the medial pterygoid medial muscle, masseter muscle and temporal muscle were cut out from their primary anatomical position and resutured to the prosthesis during the surgery and pterygoid lateral muscle have now minimal influence in distribution of stress.

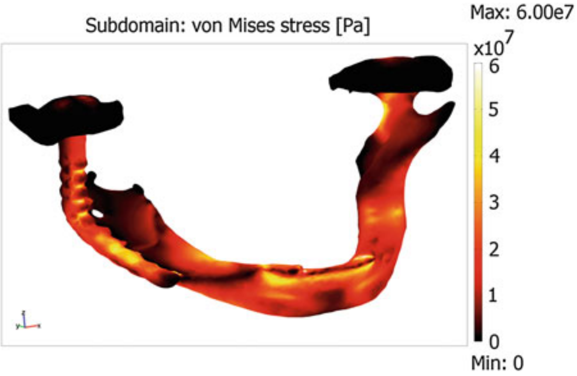


Fig. 25.7 The von Mises stress

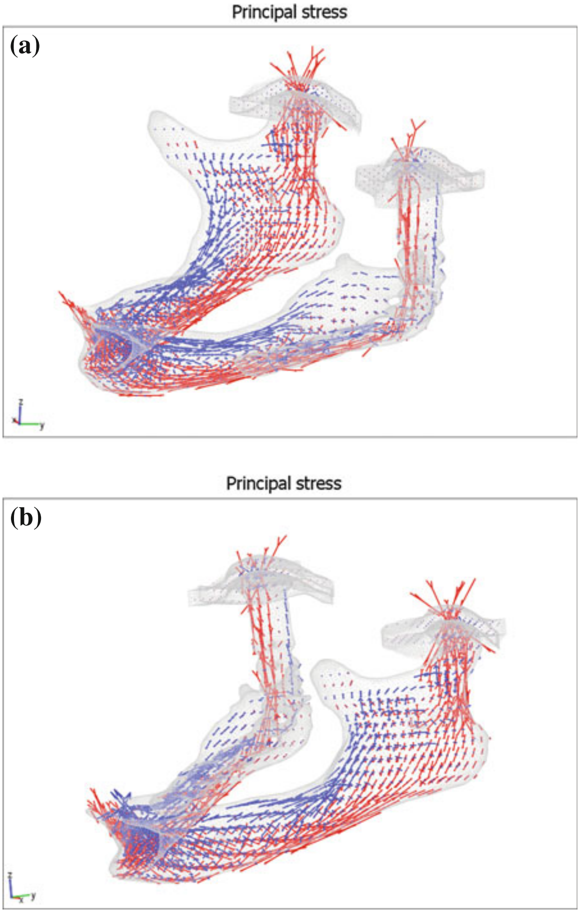


Fig. 25.8 The principal stresses: **a** back view from lower left, **b** front view from upper left

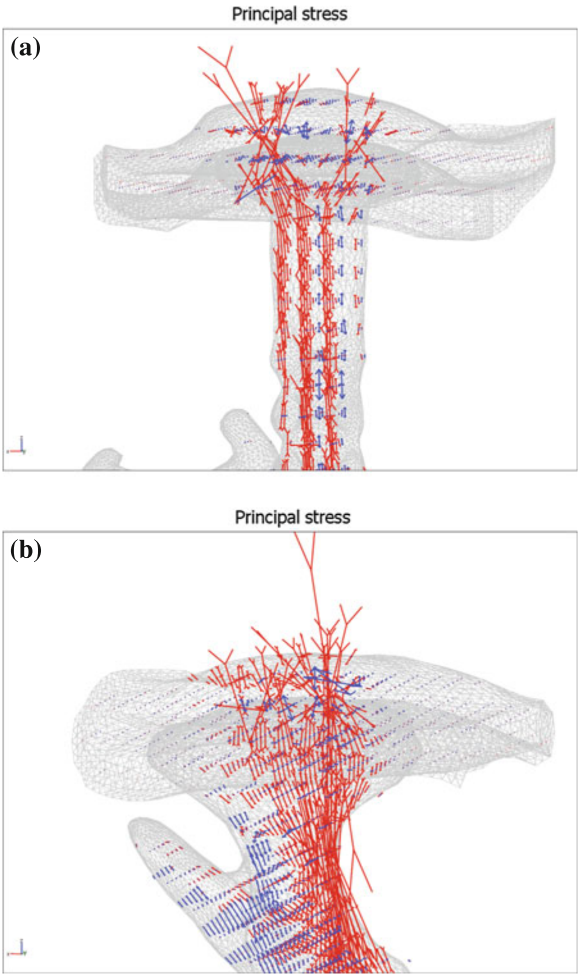
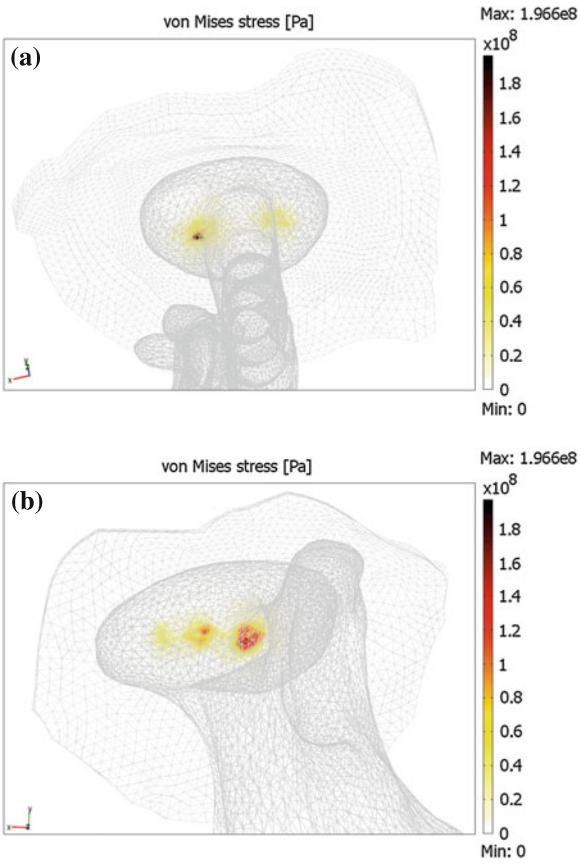


Fig. 25.9 The principal stresses—the details in the area of both condyles: **a** on the *right side* (with replacement), **b** on the *left* (healthy) side

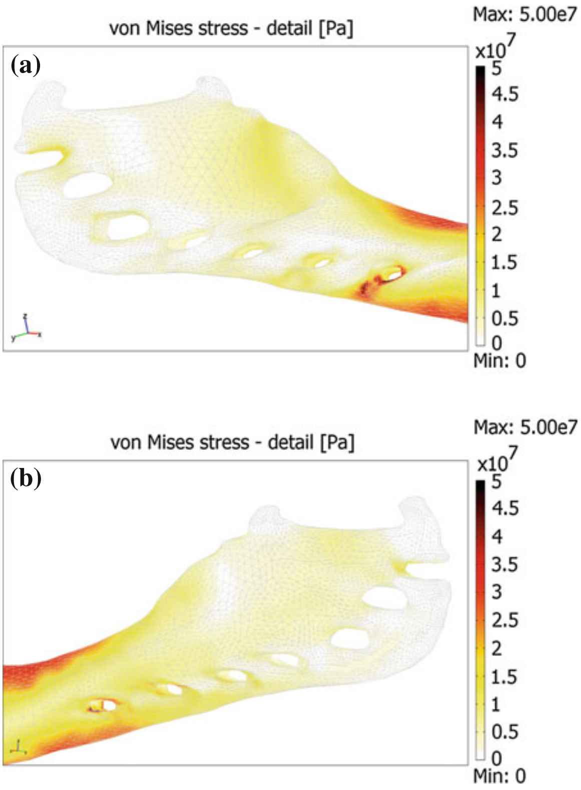
We see that the maximal values of pressures are observed in the area of the condyles of TMJ, while tensions can be observed in cranial part of corpus mandible. Numerical results show that the maximal pressure is approx. -1.05427×10^8 Pa and the maximal tension is approx. 2.02357×10^8 Pa. Interesting is the change of pressures and tensions in mandible ramus and ascendant part of the prosthesis. In the healthy left ramus are the pressures localized on posterior part and tensions on the anterior part on the right side is the opposite suitace (pressures localized anteriror and tensions posterior). It is because of changes of position of the muscle during the surgery. The studies of the detailed areas are at Figs. 25.9a,

Fig. 25.10 The von Mises stress—the details in the area of both glenoid fossa: **a** on the *right side* (with replacement), **b** on the *left* (healthy) side



b, 25.10a, b and 25.11a, b. At Fig. 25.9a, b the principal stresses are presented while at Figs. 25.10a, b and 25.11a, b the von Mises stresses are presented. The von Mises stresses shows that the loading of the glenoid fossa (acetabulum) is spread evently in the healthy TMJ, where the stresses have lower values then in the TMJP case and the stresses is situated into three greater areas, while in the TMJP case the stresses is accumulated into two areas with maximal stresses internal part of the head of TMJP. The principal stresses show how the loads are transferred into the mandible and which parts of the mandible are pressed and/or are strained by bending. The most loaded part of the area of bone and prosthesis connection is outer neighbourhood of hole for first screw especially (see Fig. 25.11a, b).

Fig. 25.11 The von Mises stress—the details in the area of holes for screws: **a** view form outside, **b** view from inside



Acknowledgments The work was partially supported by the European Regional Development Fund (ERDF), project "NTIS - New Technologies for the Information Society", European Centre of Excellence, CZ.1.05/1.1.00/02.0090, by the Grant IGA MZCR NT/13351-4 and by the long-term strategic development financing of the Institute of Computer Science (RVO:67985807).

References

1. Browne RM (1990) Investigative pathology of odontogenic cysts, 1st edn. CRC Press, Boca Raton
2. Daněk J, Dostálová T, Hubáček M, Mahdian N, Nedoma J (2012) Stress-strain analysis of the temporomandibular joint with subtotal prosthesis. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science (2012) WCECS 2012, 24–26 Oct 2012. San Francisco, pp 150–155
3. Daněk J, Hliňáková P, Přečková P, Dostálová T, Nedoma J, Nagy M (2010) Modelling of the temporomandibular joints and role of medical informatics in stomatology. Lect Notes Comput Sci 6019:62–71
4. Hlaváček I, Nedoma J (2002) On a solution of a generalized semi-coercive contact problem in thermo-elasticity. Math Comput Simul 60:1–17. doi:10.1016/S0378-4754(01)00433-5

5. Hliňáková P, Dostálová T, Daněk J, Nedoma J (2009) Temporomandibular joint and its two-dimensional and threedimensional modelling. Math Comput Simul 80(6):1256–1268. doi:10.1016/j.matcom.2009.08.007
6. May B, Saha S, Saltzman M (2001) A three-dimensional mathematical model of temporomandibular joint loading. Clin Biomech 16:489–495
7. Nedoma J (2012) Mathematical models of odontogenic cysts and of fractures of jaw-bones. An introductory study. TR-1166, ICS ASCR, Prague
8. Nedoma J, Stehlík J, Hlaváček I, Daněk J, Dostálová T, Přečková P (2011) Mathematical and computational methods in biomechanics of human skeletal systems. An Introduction. Wiley, Hoboken
9. Nishio C, Tauinoto K, Hirose M, Horiushi S, Kuroda S, Tanne K, Tanaka E (2009) Stress analysis in the mandibular condyle during prolonged clenching: a theoretical approach with the finite element method. J Eng Med 223(6):739–748. doi:10.1243/09544119JEIM485
10. Shear M, Speight PM (2007) Cysts of the Oral and Maxillofacial Regions, 4th edn. Blackwell Publishing Professional, Oxford. ISBN 978-14051-4937-2
11. Ward JP, Magar V, Franks SJ, Landini G (2004) A mathematical model of the dynamics of odontogenic cyst growth. Anal Quant Cytol Histol 26(1):39–46
12. Weijs WA, Hillen B (1985) Cross-sectional areas and estimated intrinsic strength of the human jaw muscles. Acta Morphol Nerrl Scand 23:267–274
13. Wolford LM, Mehra P (2000) Custom-made total joint prostheses for temporomandibular joint reconstruction. Proc (Bay Univ Med Cent) 13(2):135–138

Chapter 26
Fault-Tolerant Optimization for Application-Specific Network-on-Chip Architecture

Farnoosh Hosseinzadeh, Nader Bagherzadeh, Ahmad Khademzadeh and Majid Janidarmian

Abstract Advanced integration technologies enable the construction of Network-on-Chip (NoC) from two dimensions to three dimensions. Studies have shown that 3D NoCs can improve average communication performance because of the possibility of using the additional dimension to shorten communication distance. This paper presents a defect tolerance technique for recovering permanent routers failure through an efficient and effective use of redundancy. This technique is ideally suited for three and even two dimensional network-on-chip (NoC). This fault-tolerant NoC architecture designed in VHDL and synthesized using Xilinx ISE is presented. Simulation results demonstrate significant reliability and yield improvement. Although the hardware overhead of the 3D (2D)-proposed methodology compare with traditional mesh is approximately 15 % (12 %), it improves the average response time of system up to 31 % (23 %).

Keywords Application-specific · By-pass · Defect tolerant architecture · Mapping · Network-on- chip · Spare router

F. Hosseinzadeh (✉) · M. Janidarmian
Science and Research Branch, Islamic Azad University, Tehran, Iran
e-mail: fh.zadeh@srbiau.ac.ir

M. Janidarmian
e-mail: jani@ieee.org

N. Bagherzadeh
University of California, Irvine, USA
e-mail: nader@uci.edu

A. Khademzadeh
Iran Telecommunication Research Center, Tehran, Iran
e-mail: zadeh@itrc.ac.ir

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_26,

26.1 Introduction

The fast developing integrated circuit (IC) manufacturing technology has provided the industry with billions of transistors on a single chip [1, 2]. At the same time, the numbers of circuits integrated on a chip have been increasing which leads to an exponential rise in the complexity of their interaction. Traditional digital system design methods, e.g., bus-based System-on-Chip (SoC) will encounter performance bottlenecks. One of the most well known problems is the communication bottleneck. Most SoCs have bus-based communication architectures, such as simple, hierarchical or crossbar-type buses. Bus based systems do not scale well with the system size in terms of bandwidth, clock frequency and power consumption [3]. To address these problems, NoC was proposed as a promising communication platform solution for future multicore systems [3].

On-chip networks will be prevalent in computing domains ranging from high-end servers to embedded system-on-chip (SoC) devices. This diversity of application platforms has led to research in on-chip networks spanning a variety of disciplines from computer architecture to computer-aided design, embedded systems, VLSI and more [4].

Besides design and verification benefits, NoCs have also been advocated to address increasingly daunting clocking, signal integrity, and wire delay challenges. Indeed, tremendous progress has been made in recent years on the design of 2D NoC architectures, both on regular topologies like 2D mesh networks for chip-multiprocessor applications [5–8] and on application specific network architectures for custom SoC designs [9–12]. However, the advent and increasing viability of 3D silicon integration technology have opened a new horizon for new on-chip interconnects design innovations.

3DNoCs overcome the limited scalability of 2DNoCs over 2D planes by using short and fast vertical interconnects of 3D-ICs. Compared with 2DNoCs, 3DNoCs greatly reduce the network diameter and overall communication distance, thus improving communication performance and reducing power consumption. Till today extensive results have shown that 3D networks improve 2D network scalability [13] and performance in terms of delay and throughput [14, 15].

However, a major challenge facing the design of such highly integrated 3D-ICs in deep submicron technologies is the increased likelihood of failure due to permanent and intermittent faults caused by a variety of factors that are becoming more and more prevalent. Permanent faults occur due to manufacturing defects, or after irreversible wear out damage due to electro migration in conductors, negative bias temperature instability, dielectric breakdown, etc., [16, 17]. Intermittent faults on the other hand, occur frequently and irregularly for several cycles, and then disappear for a period of time [18, 19]. These faults commonly arise due to process variations combined with variation in the operating conditions, such as voltage and temperature fluctuations. This argument strengthens the notion that chips need to be designed with some level of built-in fault tolerance. Furthermore, relaxing of 100 % correctness in the operation of various components and channels profoundly reduces the manufacturing cost as well as cost incurred by test and verification [20].

This paper is an extension of [21] and the rest of this paper is organized as follows. In Sect. 26.2, an over view of some fault-tolerant research efforts in NoC is given. Section 26.3 illustrates the basic concepts of application-specific NoC design, and a new fault-tolerant architecture is introduced in Sect. 26.4. The results are given in Sect. 26.5 followed by the concluding remark in Sect. 26.6.

26.2 Related Work

Three-dimensional integrated circuits (3D ICs) [22] offer an attractive solution for overcoming the barriers to interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology, with smaller form factor, higher integration density, and the support for the realization of mixed-technology chips. Among several 3D integration technologies [23], TSV (Through-Silicon-Via) approach is the most promising one and therefore is the focus of the majority of 3D integration R&D activities [22]. Even though both 3D integrated circuits and NoCs [24, 25] are proposed as alternatives for the interconnect scaling demands, the challenges of combining both approaches to design three-dimensional NOCs have not been addressed until recently [26].

Pavlidis and Friedman [27] have compared 2D mesh NoC with its 3D counterpart by analyzing the zero-load latency and power consumption of each network. In the work of [28], a performance analysis method based on network calculus has been proposed for 3DNoC. In [29], the performance of several alternative vertical interconnection topologies has been studied. In [30], the authors proposed a dimension decomposition scheme to optimize the cost of 3DNoC switches, and presented some area and frequency figures derived from a physical implementation.

Fault-tolerant design is a design that enables a system to continue operation, possibly at a reduced level, rather than failing completely, when some part of the system fails.

Fault-tolerant routing algorithms should be able to find a path from source to destination in presence of the faults in NoC with a certain degree of tolerance [31]. Many algorithms in this area have been proposed which follow their own optimization aims.

In [32], presents the scalable and fault-tolerant distributed routing (SFDR) mechanism. It supports three routing modes including corner-chains routing, boundary-chains routing and fault-ring routing.

A fault-tolerant mesh-based NoC architecture with the ability of recovering from single permanent failure is presented in [33]. This method adds a redundant link between each core and one of its neighboring routers, resulting in significant improvement in reliability while has little impact on performance. In [34], a hardware and performance aware design for the fault-tolerant 2DNoC architecture is presented.

In the 3D domain, [35] present an application specific 3DNoC synthesis algorithm that is based on a rip up-and-reroute procedure for routing flows, where the traffic flows are ordered in the order of increasing rate requirements so that smaller flows are routed first, followed by a router merging procedure. Murali et al. [36] propose a 3D NoC topology synthesis algorithm, which is an extension to their previous 2D work [37].

In this paper, a new fault-tolerant application-specific network-on-chip was proposed which is able to tolerate routers failure and guarantees the 100 % packet delivery.

26.3 Perquisites of Application-Specific NoC Design

26.3.1 Mapping Problem

To formulate mapping problem in a more formal way, we need to first introduce the following two concepts borrowed from [38].

Definition 1 The core graph is a directional graph G (V, E), whose each vertex $v_i \in V$ shows a core, and a directional edge $e_{i,j} \in$ E illustrates connection between v_i and v_j The weight of $e_{i,j}$ that is shown as $comm_{i,j}$, represents the communication volume from v_i to v_j. The IP core along with a router connected to it by Network Interface (NI) is displayed as a tile. The definition is presented in Fig. 26.1.

Definition 2 The NoC architecture graph is a directional graph $A(T, L)$, whose each vertex $t_i \in T$ represents a tile in the NoC architecture, and its directional edge that is shown by $l_{i,j} \in L$ shows a physical link from t_i to t_j.

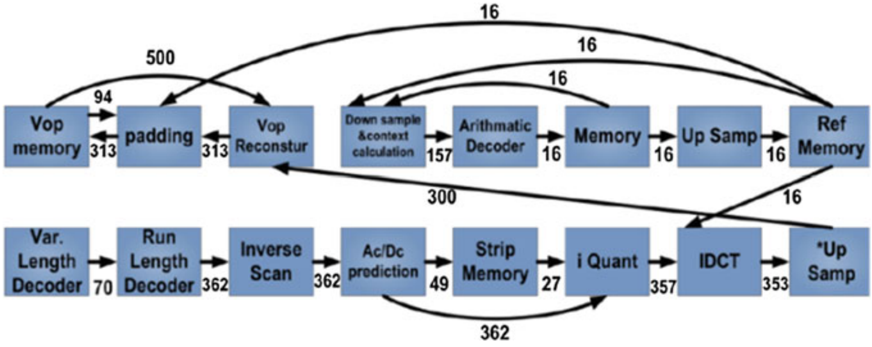


Fig. 26.1 The VOPD core graph

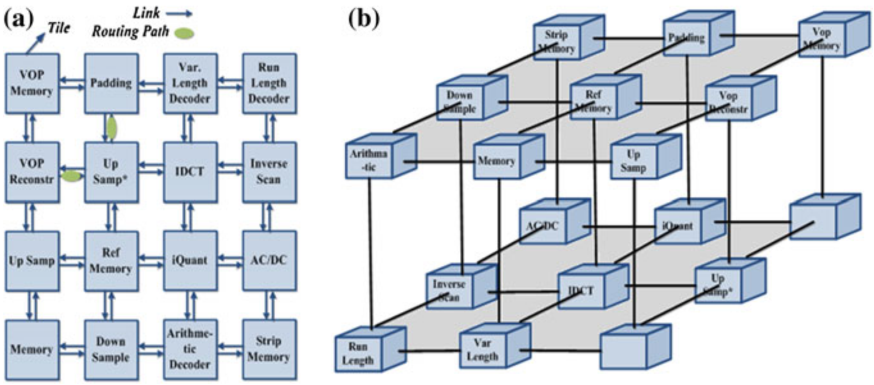


Fig. 26.2 Mapping of VOPD core graph. **a** 2DMesh. **b** 3DMesh

The core graph mapping $G(V, E)$ on NoC architecture graph $A(T, L)$ is defined by a one to one mapping function (26.1):

$$map : V \rightarrow T, \; s.t.map(v_i) = t_j, \; \forall v_i \in V, \; \exists t_j \in T, \; |V| \leq |T| \tag{26.1}$$

Mapping algorithms are mostly focused on mesh topology (Architecture Graph) which is the most popular topology in NoC design due to its suitability for on-chip implementation and low cost. Mapping of VOPD Core graph on 2D and 3Dmesh is displayed in Fig. 26.2.

26.3.2 Routing Algorithm

The routing algorithm determines the path that each packet follows between a source–destination pair. Routing algorithms noticeably affect the cost and performance of NoC parameters i.e., area, power consumption and average message latency. Due to determined sources and destinations in application-specific NoC minimum-distance routing algorithms are mostly considered in this area which is computed off-line and admissible paths stored into the routing tables. For avoiding any possible deadlocks, we used the concept of application specific channel dependency graph (ASCDG) introduced in [39].

The proposed methodology is topology and application agnostic, the state-of-art mapping algorithm proposed in [40] is used to map Video Object Plan Decoder (VOPD) cores onto mesh topology. We have used minimum-distance routing algorithm in 2D(3D)mesh i.e., XY(XYZ). In XY(XYZ) routing algorithm, a packet is routed first along the X axis, then the Y axis (and finally the Z axis).

26.3.3 Average Response Time and Communication Cost

The average response time of the system are considered as the evaluation parameters. Using (26.2), one can evaluate delay for sending one bit on communication channel from node i to node j in mesh topology:

$$T_{bit}^{n_i,n_j} = \sum_{k=1}^{n\,hops} (T_{S_k bit} + T_{fifo_k bit}) + (n_{hops} - 1) \times T_{Lbit} \tag{26.2}$$

$T_{bit}^{n_i,n_j}$ indicates transmission delay of one bit between two nodes $T_{S_k bit}$ and T_{Lbit} represent forwarding delay of one bit in switch and link, respectively. n_{hops} is waiting time in the input FIFO of the switches. Delay for transmitted data could be calculated by (26.3)

$$T_{bit}^{n_i,n_i} = Data\,size \times T_{bit}^{n_i,n_i} \tag{26.3}$$

In this simulation model, the weight of ei, j that is shown as commi, j, represents the volume of the communication from vi to vj. Consequently, the average response time of the system is estimated by maximum delays of last bits arriving at the destination cores.

The communication cost is calculated by (26.4)

$$Comm\cos t = \sum_{k=1}^{|E|} (V_L(d^k) \times dist(Source(d^k), Dest(d^k))) \tag{26.4}$$

The communication cost is calculated when all switches work correctly.

If a switch fails, communication cost is increased because of some other routing paths.

26.4 The Proposed Fault-Tolerant Design

In the mesh-based architecture which is the simplest and most dominant topology for today's regular 2D(3D) network on chips, each core is connected to a single router.

The natural and simplest extension to the baseline NoC router to facilitate a 3D layout is simply adding two additional physical ports to each router; one for Up and one for Down, along with the associated buffers, arbiters (VC arbiters and Switch Arbiters), and crossbar extension. We can extend a traditional NoC fabric to the third dimension by simply adding such routers at each layer.

If a failure occurs in a router in this topology, the failed router cannot be used any more for routing packets and the directly connected core obviously loses its communication with the network, so the expected requirements of the mapped application are not satisfied and the whole system breaks down.

In order to recover the inaccessibility of the core, we provide two steps:

(1) Applying the spare router and new link interface. (2) Bypassing ports in each router to increase reliability and performance.

26.4.1 Selecting Spare Router and Using Link Interface

By assuming a faulty router in a mesh-based NoC, it is obvious that the core directly connected to the faulty router is inaccessible. Hence, each core is connected two routers. They are main and spare routers which have been proposed [33]. A router can inform the neighbors about its faulty status by setting a fault-status flag. This flag is checked by the neighboring routers and cores before starting any communication with the router.

Adding one port to all none-edge routers is not desirable so to minimize hardware overhead, we used a component called Link Interface instead of router port. This module is suggested in order to reduce the overhead. After entrance of header flit into this module, destination address is decoded. As an example, if the address shows that connected core is the destination, header and its following flits will be sent to backup network interface of the core, otherwise they are routed to another output towards next router.

The Link Interface has been implemented with three processes which run concurrently; therefore it is able to transmit three dataflow as shown in Fig. 26.3. This module has been also designed without using clock pulse that leads us to achieve better response time and power consumption.

Furthermore, we modify algorithm in [41] not only for given the best result in average response time, extra communication cost and system reliability, but also for having more routing path opportunities. This algorithm explained in details in Fig. 26.4.

You can see in Figs. 26.5, 26.6 the proposed architecture which implies in this step. Each core is connected to its router via main local port and to the links using Link Interface via backup local port.

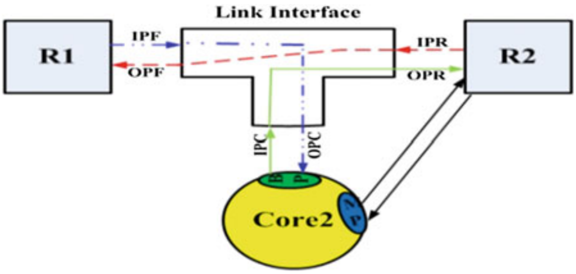
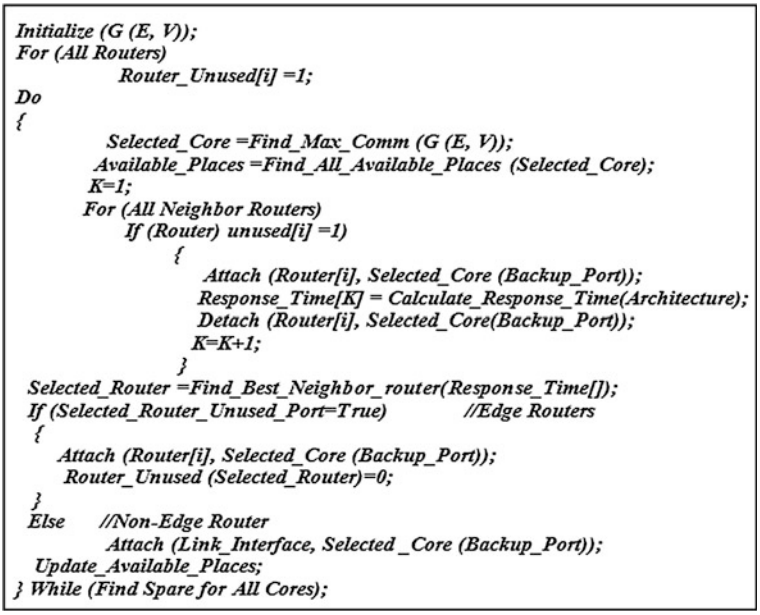


Fig. 26.3 The module link interface



```
Initialize (G (E, V));
For (All Routers)
          Router_Unused[i] =1;
Do
{
         Selected_Core =Find_Max_Comm (G (E, V));
        Available_Places =Find_All_Available_Places (Selected_Core);
       K=1;
       For (All Neighbor Routers)
            If (Router) unused[i] =1)
                 {
                   Attach (Router[i], Selected_Core (Backup_Port));
                   Response_Time[K] = Calculate_Response_Time(Architecture);
                   Detach (Router[i], Selected_Core(Backup_Port));
                  K=K+1;
                 }
 Selected_Router =Find_Best_Neighbor_router(Response_Time[]);
 If (Selected_Router_Unused_Port=True)          //Edge Routers
  {
    Attach (Router[i], Selected_Core (Backup_Port));
     Router_Unused (Selected_Router)=0;
  }
 Else    //Non-Edge Router
         Attach (Link_Interface, Selected _Core (Backup_Port));
  Update_Available_Places;
} While (Find Spare for All Cores);
```

Fig. 26.4 The pseudo code of selecting the best spare router

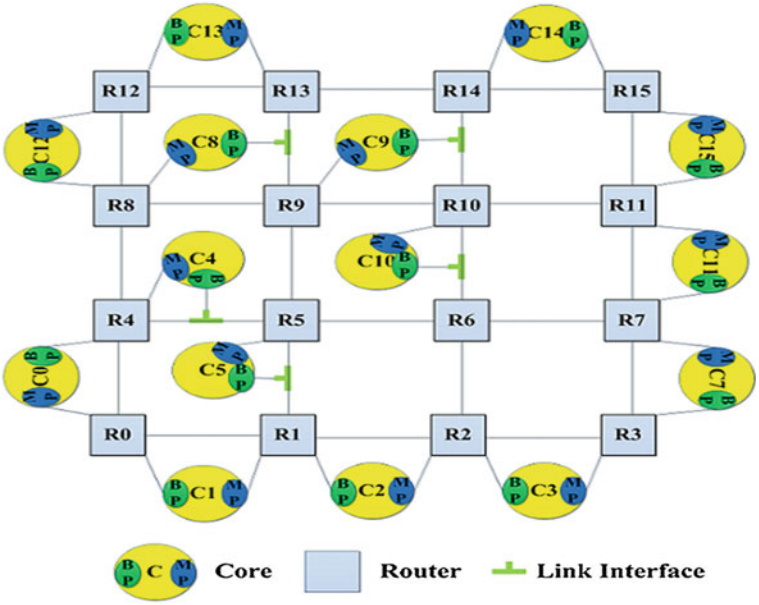


Fig. 26.5 The module link interface on 2DMesh

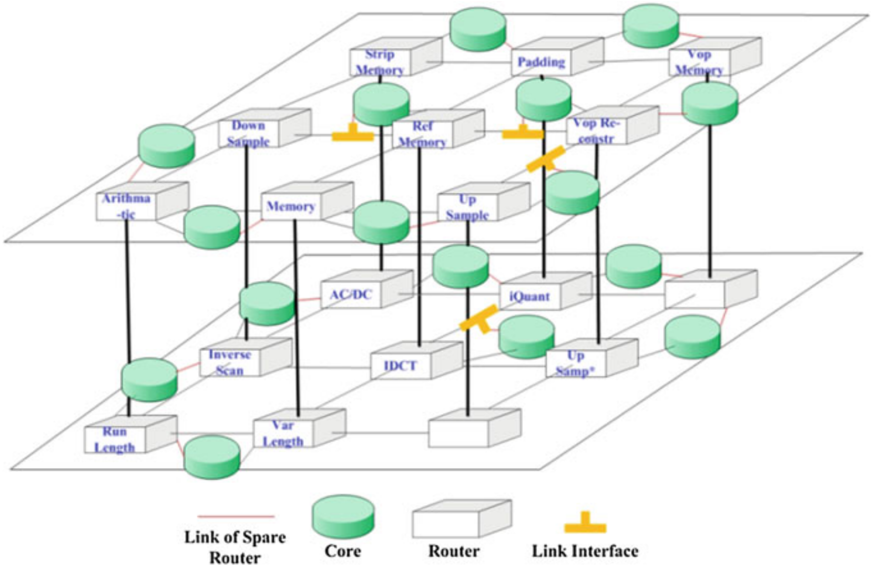


Fig. 26.6 The module link interface on 3DMesh

26.4.2 Bypassing Ports in a Router Based on Average Response Time of the System

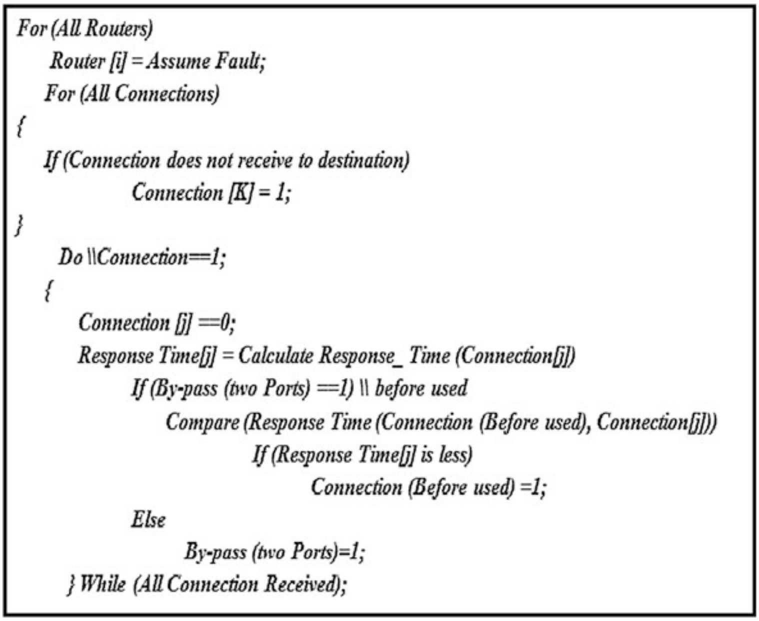
When a router is failed, the packet reaches to destination by through its spare router. However, if special routing such as XY (XYZ) which we apply for our routing is used, some packets will not arrive to their destinations.

In a failed router two ports are bypassed to each other until a packet arrive to destination. There is constraint to select ports: each port is limited only one other port in the router. At the beginning, when a router fails, are indicated the connections do not reach to their destinations. Then the two ports which not only can arrive a packet to its destination, but also average response time is least are selected. This algorithm is discussed in details in Fig. 26.7.

Table 26.1 indicates the best ports are bypassed in each router in 2D(3D)mesh that is the left section for 2D mesh and the right section for 3D one, and the abbreviations shows N: North Port, E: East Port, W: West Port, S: South Port, U: UP Port and D: Down Port.

26.5 Experimental Results

Fault tolerance evaluates how reliability architecture can rout messages in despite different types of faults. Video Object Plan Decoder (VOPD) as a case study has been mapped on a 4 × 4(3 × 3 × 2) mesh topology using the best mapping



```
For (All Routers)
   Router [i] = Assume Fault;
   For (All Connections)
{
   If (Connection does not receive to destination)
           Connection [K] = 1;
}
    Do \\Connection==1;
   {
      Connection [j] ==0;
      Response Time[j] = Calculate Response_ Time (Connection[j])
           If (By-pass (two Ports) ==1) \\ before used
              Compare (Response Time (Connection (Before used), Connection[j]))
                     If (Response Time[j] is less)
                            Connection (Before used) =1;
           Else
                By-pass (two Ports)=1;
     } While (All Connection Received);
```

Fig. 26.7 The pseudo code of bypassing ports

Table 26.1 Selecting the best bypassing ports in each router

0		1		2		3	
L-E	E-N L-W	W-E	E-L	E-N	N-E	–	D-L
4		5		6		7	
S-L	–	W–S L-N	E-N L-D	S-E L-N	U-L	N-L	L-U S-E
8		9		10		11	
N-L	N-L	S–N	L-U	–	U-L	N-L	S-L
12		13		14		15	
E-L	W-L	S-L	W-D S-L	–	–	W-L	E-L

technique proposed in [38]. To evaluate our design, we perform some most important fault tolerance metrics: reliability, MTTF, yield, response time, communication cost and hardware cost.

26.5.1 Reliability Analysis

The reliability of an NoC router $R_r(t)$ is the probability that a router performs its functionalities correctly from time 0 to time t. It is decided by the failure rate, $\lambda(t)$, which is measured by the number of failures per time unit. After the router has passed the infant mortality period, we can express $R_r(t)$ using exponential failure law [42]:

$$R_r(t) = e^{-\lambda t} \tag{26.5}$$

To calculate the reliability of the system, simple serial and parallel rules are used; hence for each pair of source and destination nodes in core graph, the reliability of the path is production of reliability of the routers which are met through the unique path according to XY(XYZ) algorithm. Then the system reliability is calculated again, each time assuming each individual router fails and accordingly a new rerouting path is contemplating XY(XYZ) algorithm and the additional spare routers. The final analytical formulation for system reliability is as (26.6):

$$\begin{gathered} for\,3D \quad l = (i, j, k) \\ for\,2D \quad l = (i, j) \\ R_{NOC} = \prod\nolimits_{(l) \in CTG} \left[\coprod\nolimits_{t=1}^{s} R_t + \sum\nolimits_{t=1}^{s} (1 - R_t) \cdot (R_{l)|Router_t Fails}) \right] \end{gathered} \tag{26.6}$$

In which, (i, j, k) represent each pair of source and destination nodes in core graph. “S” shows the number of routers in each routing path and $R_{(i,j) or (i,j,k)}$ is the reliability of the path from i to j (and then to k). We assume the failure rate of a router is $\lambda = 0.00315$ (times/year) [43].

Furthermore, for showing effect of mapping and number of TSVs that is used in sending packets in 3D architecture, we execute different algorithm mappings in VOPD application such as Onyx [26], using with least TSVs (LTSV), and using with most TSVs (MTSV). The reliability of 3Dmesh compare to our design over 1–50 years is shown in Fig. 26.8.

Moreover, the reliability of the links is affected in reliability of system what is as (26.7). The failure rate of a horizontal link is $\lambda_{LH} = 0.0088$ (times/year) [43]. Because probability failure of vertical links is more than horizontal links [27], we assume the failure rate of a vertical link is $\lambda_{LV} =$ N. λ_{LH} (N = 1, 3, 6, 12).

$$R_{NOC} = \prod\nolimits_{(i,j,k) \in CTG} \left[\prod\nolimits_{t=1}^{s} R_t . R_{Link|LinkTSV} + \sum\nolimits_{t=1}^{s} (1 - R_t) . (R_{(i,j,k).} R_{Link|LinkTSV} | Router_t Fails \right] \tag{26.7}$$

The reliability of 3Dmesh compare to our design when add reliability of links is displayed in Fig. 26.9, also the probability faulty free connections when one to three routers are failed are displayed in Table 26.2.

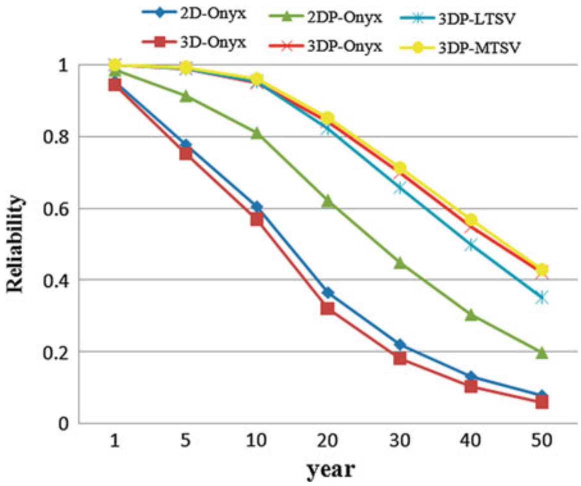


Fig. 26.8 Reliability of mesh and our design

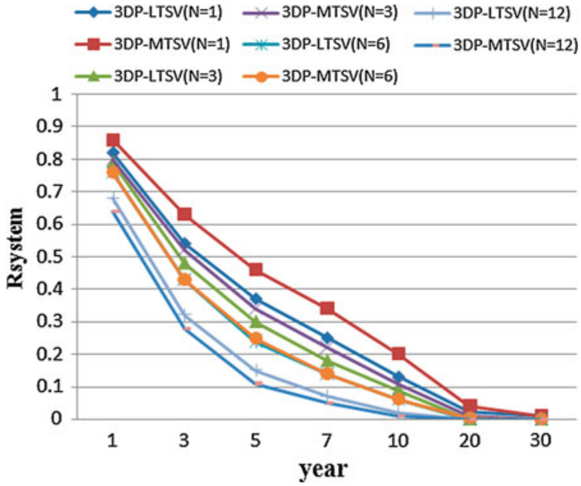


Fig. 26.9 Reliability of our proposed design in two different mappings

Table 26.2 Probability faulty-free connections

3DNoC (%)	2DNoC (%)	Number of faulty routers
100	100	1
92	87	2
78	68	3

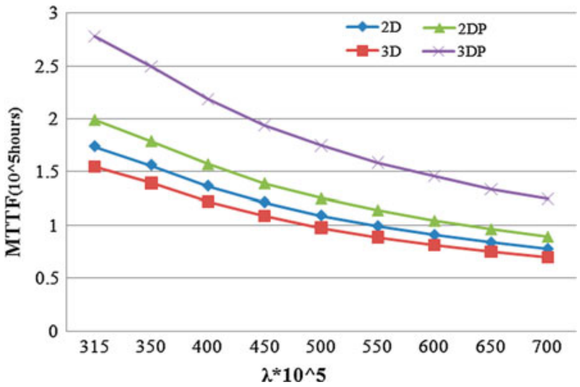


Fig. 26.10 Mean Time to Failure of 2Dmesh and our design

26.5.2 Mean Time to Failure Analysis

The mean time to failure (MTTF) is the average time before a system fails which can be expressed as the area under the reliability curve [19]:

$$MTTF = \int_0^{\infty} R(t)dt \tag{26.8}$$

The MTTF for R^N can be derived as the following:

$$MTTF_{R^n} = \int_0^{\infty} e^{-n\lambda t}dt = \frac{1}{n * \lambda} \tag{26.9}$$

The MTTF for mesh and our spared mesh is shown in Fig. 26.10.

26.5.3 Yield Analysis

The yield of NoC is modeled by the yield of block and the yield of connection. Recent research has shown that correlation factor (Ω) between components should also be introduced in the yield calculation [44]. According to [44], our system's yield is modeled by Eq. (26.10).

$$y_{NoC} = (y_{Block})^n \cdot y_{conn} \tag{26.10}$$

We assume a channel has 64-bit data wires, 4-bit control and 8-bit parity wires. The parameters used in our yield evaluation work are listed in Table 26.3 [43].

We perform the Monte Carlo simulation for the NoC yield. Figure 26.11 shows the flow chart of our simulation. The experimental results are shown in Fig. 26.12. The yield of 3DNoC is very low [27], but the yield of system has improved 17 % with applying our fault tolerant architecture, also has enhanced 14 % in 2D one.

Table 26.3 Parameters for yield calculations

α	Clustering parameter	2	y_i	True yield of other components	99.5 %
Ω_i	Defect coverage for all components	99 %	y_{w-ctrl}	Yield of control or parity wires	99.99 %
y_c	True core yield	97 %	σ_{i-j}	Correlation among two components	0.5

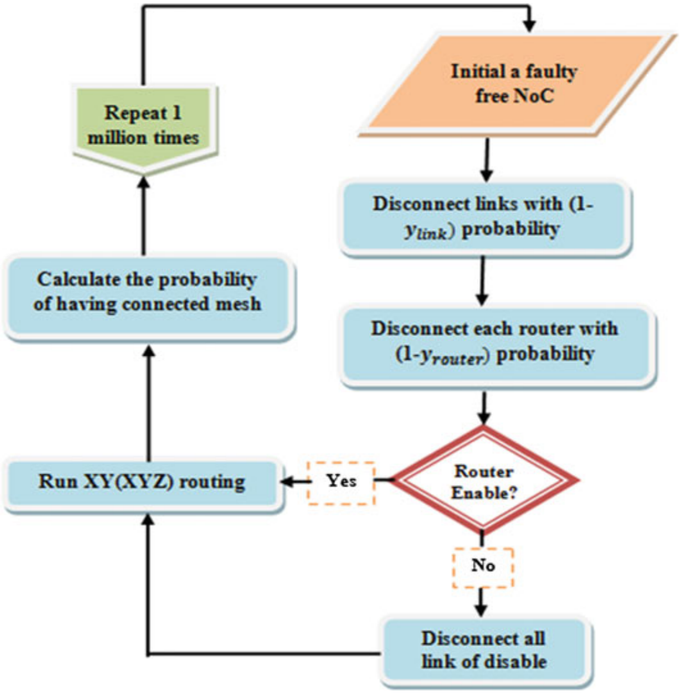
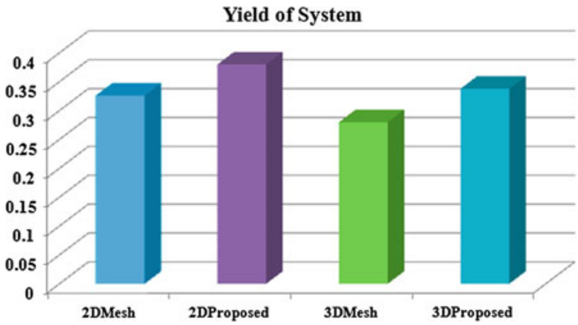

Initial a faulty free NoC
Disconnect links with (1-y_{link}) probability
Disconnect each router with (1-y_{router}) probability
Router Enable?
Yes
No
Disconnect all link of disable
Run XY(XYZ) routing
Calculate the probability of having connected mesh
Repeat 1 million times

Fig. 26.11 Monte Carlo simulation

Fig. 26.12 The yield analysis



26.5.4 Average Response Time and Communication Cost Analysis

In order to compare the average response time of the reliable architecture and 2D (3D) mesh, they have been designed in VHDL and synthesized using Xilinx ISE. The proposed architecture significantly decrease the average response time on the faultless and 16(18) possible faulty routers compared with the 2D (3D)mesh architecture as illustrated in Table 26.4. Also, considering the results of mathematical calculation of communication cost is shown in Fig. 26.13.

It should be pointed that it actually is a great achievement to develop a fault-tolerant NoC design which also has better performance. To explain in details, when all routers are correctly operating, new architecture improves the average response time by 31 % (23 %) comparing to 3D (2D)mesh.

In Fig. 26.14 we observe that the proposed approach allows decreasing the response time of system by 21 % (16 %) and tolerating permanent failure of each single router.

26.5.5 Hardware Cost

To recover from a permanent fault, hardware redundancy is mandatory and reduction of this overhead has always been an important issue in this area. In our design, we do not add any router port and instead a link interface was developed which helps to achieve less hardware overhead.

The proposed architecture designed and implemented in the Vertex E FPGA (v50ecs144-6). Figure 26.15 indicates that our proposed fault-tolerant architecture added 12 % (15 %) redundancy camper with 2D (3D)Mesh in the worst case.

Table 26.4 Average response time of system in different situations

R	N = 2	N = 3	R	N = 2	N = 3	R	N = 2	N = 3
R0	39.08	36.77	R1	38.01	33.21	R2	38.01	28.8
R3	41.83	32.16	R4	40.15	32.64	R5	43.95	28.97
R6	44.1	34.06	R7	43.27	32.25	R8	38.01	28.8
R9	41.59	37.2	R10	39.85	32.21	R11	40.15	32.49
R12	38.97	39.66	R13	42.09	39.66	R14	41.59	32.21
R15	40.03	34.54	R16	–	32.16	R17	–	32.3
NO-FAULT	37.26	28.8	N-Mesh	48.2	41.8			

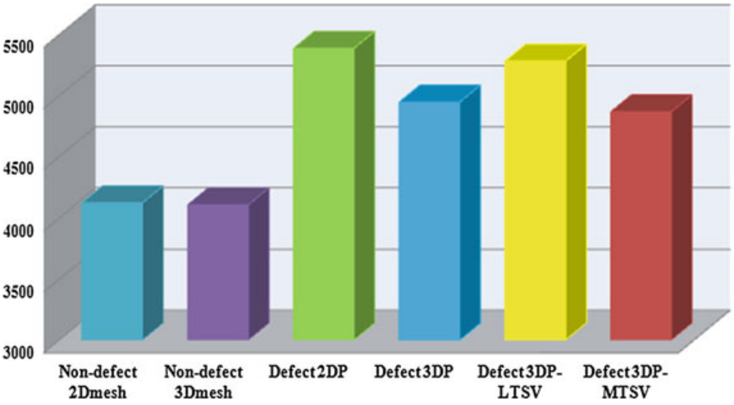


Fig. 26.13 Communication cost in different mappings

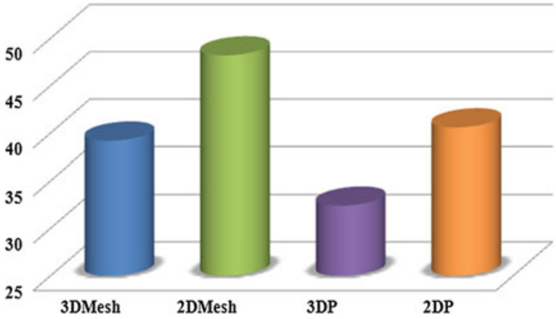


Fig. 26.14 Comparing proposed architecture with Mesh in terms of average response time

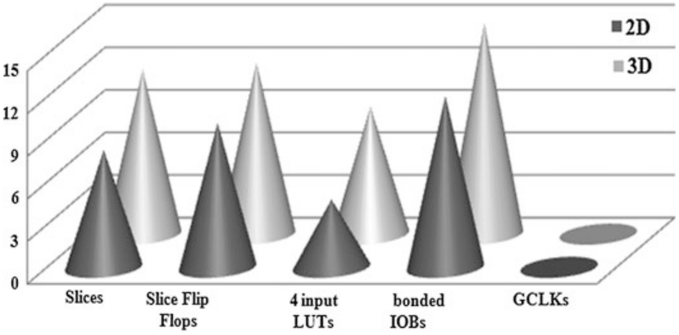


Fig. 26.15 Comparing hardware overhead proposed design with 3DMesh (all utilities compare with percent)

26.6 Conclusion

To achieve the targeted reliability, the errors should be recovered. Permanent error recovery results in huge area and energy overhead. In this paper, a fault-tolerant application-specific architecture has been proposed to improve latency and yield with up to 15 % area overhead in 3DNoC. Also, we have shown that 2D NoC with this design can be increased the reliability of system. This architecture is topology, application, mapping agnostic, but using a mapping with at least hop count is very important, because it leads to improving reliability, MTTF, and yield. Furthermore, this fault tolerant approach designed in VHDL and synthesized using Xilinx ISE. Simulation results show that the proposed method in 3D (2D) NOC reduces latency by 31 % (23 %) and achieves an up to 17 % (14 %) higher yield, compared to base Mesh. In 3DMesh, we must apply a mapping with using at least TSV and Extra Bandwidth. This leads to improving reliability, and even we take advantage of spare TSVs.

References

1. Intel (2008) Intel core i7 processor extreme edition and intel core i7 processor datasheet, vol 1
2. AMD (2010) The amdopteron 6000 series platform http://www.amd.com/us/products/server/processors/6000-seriesplatform/pages/6000-series-platform.aspx May2010
3. Dally WJ, Towles B (2001) Route packets, not wires: on-chip interconnection networks. In: Proceedings of the 38th conference on Design automation, June 2001, pp 684–689
4. Jerger NE, Peh L-S (2009) On-chip networks, Synthesis Lectures on Computer Architecture 2009
5. Taylor et al. MB (2002) The RAW microprocessor: a computational fabric for software circuits and general-purpose programs. IEEE Micro 22(6):25–35
6. Sankaralingam et al K (2003) Exploiting ILP, TLP, and DLP with the polymorphous TRIPS architecture, ISCA
7. Hu J, Marculescu R (2003) Energy-aware mapping for tile-based NoC architectures under performance constraints, ASP-DAC
8. Murali S, De Micheli G (2004) Bandwidth constrained mapping of cores onto NoC architectures
9. Srinivasan K, Chatha KS, Konjevod G (2006) Linear-programming based techniques for synthesis of network-on-chip architectures. IEEE Trans VLSI Syst, Apr
10. Murali et al. S (2006) Designing application-specific networks on chips with floor plan information, ICCAD
11. Yan S, Lin B (2008) Application-specific network-on-chip architecture synthesis based on set partitions and Steiner trees, ASPDAC
12. Yan S, Lin B (2008) Custom networks-on-chip architectures with multicast routing. IEEE Trans VLSI Syst, accepted for publication
13. Weldezion AY, Grange M, Pamunuwa D, Lu Z, Jantsch A, Weerasekera R, Tenhunen H (2009) Scalability of network-on-chip communication architecture for 3D meshes. In: Proceedings of the 3rd ACM/IEEE international Symposium on networks-on-chip (NOCS'09), May 2009
14. Feero B, Pande P (2008) Networks-on-chip in a three-dimensional environment: a performance evaluation. IEEE Trans Comput

15. Kim J, Nicopoulos C, Park D, Das R, Xie Y, Vijaykrishnan N, Yousif MS, Das CR (2007) A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In: Proceedings of the international Symposium on computer architecture (ISCA'07), June 2007
16. Bolotin E, Cidon I, Ginosar R, Kolodny A (2004) Cost considerations in network on chip. Integr VLSI J 38(1):19–42
17. Nassif S (2001) Modeling and analysis of manufacturing variations. Proc CICC
18. Pan et al S (2010) IVF, characterizing the vulnerability of microprocessor structures to intermittent faults. Proc
19. Constantinescu C (2007) Intermittent faults in VLSI circuits. Proc IEEE Workshop silicon Errors logic (SELSE)
20. Dumitraş T, Mărculescu R (2003) On-chip stochastic communication, vol 1, Design, Automation and test in Europe conference and exhibition, p 10790
21. Hosseinzadeh F, Bagherzadeh N, Khademzadeh A, Janidarmian M, Koupaei FK (2012) Improving reliability in application-specific 3D network-on-chip, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, USA, pp 196–202, 24–26 Oct 2012
22. Philip G, Christopher B, Ramm P (2008) Handbook of 3D integration. Wiley-VCH, New York
23. Davis W et al (2005) Demystifying 3D ICs: the pros and cons of going vertical. IEEE Des Test Comput 22(6):498–510
24. Jantsch A, Tenhunen H (2003) Networks on chip. Kluwer Academic Publishers, Berlin
25. De Micheli G, Benini L (2006) Networks on chips. Morgan Kaufmann, Burlington
26. Li et al. F (2006) Design and management of 3D chip multiprocessors using network-in-memory. In: Proceedings of international Symposium on computer architecture, pp 130–141, Jun 2006
27. Loi I, Angiolini F, Fujita S, Benini L (2011) Characterization and implementation of fault-tolerant vertical links for 3-D Networks-on-Chip. IEEE Trans Comput Aided Des Integr Circuits Syst 30(1)
28. Pavlidis VF, Friedman EG (2007) 3-D topologies for networks-on- chip. IEEE Trans Very Large Scale Integr (VLSI), pp 1081–1090
29. Feero B, Pande PP Performance evaluation for three-dimensional networks-on-chip. IEEE Comput Soc Annu Symp VLSI (ISVLSI'07)
30. Weldezion AY, Grange M, Weerasekera R, Tenhunen H (2009) Scalability of network-on-chip communication architecture for 3-D meshes, IEEE
31. Dumitras T, Kerner S, Marculescu R (2003) Towards on-chip fault tolerant communication. In: Proceedings of the Asia and South Pacific design automation conference
32. Shi Z, You K, Ying Y, Huang B, Zeng X, Yu Z (2010) A scalable and fault-tolerant routing algorithm for NoCs. Int Symp Circ, pp 165–168
33. Refan F, Alemzadeh H, Safari S, Prinetto P, Navabi Z (2008) Reliability in application specific mesh-based NoC architectures, on-line testing symposium, 14th IEEE International, pp 207–212, Jul 2008
34. Koupaei FK, Khademzadeh A, Janidarmian M (2011) Fault-tolerant application-specific network-on-chip, IEEE 2011
35. Yan S, Lin B (2008) Design of application-specific 3D networks-on-chip architectures. In: Proceedings of ICCD, pp 142–149
36. Murali S, Seiculescu C, Benini L, Micheli GD (2009) Synthesis of networks on chips for 3D systems on chips. In: Proceedings of ASPDAC, pp 242–247
37. Murali et al S (2006) Designing application-specific networks on chips with floor plan information. In: Proceedings of ICCAD, pp 355–362
38. Janidarmian M, Khademzadeh A, Tavanpour M (2009) Onyx: A new heuristic bandwidth-constrained mapping of cores onto tile-based Network on Chip. IEICE Electron Express 6(1):1–7
39. Palesi M, Holsmark R, Kumar S (2006) A methodology for design of application specific deadlock-free routing algorithms for NoC systems, hardware/software codesign and system

synthesis, CODES + ISSS ‘06. In: Proceedings of the 4th international conference, pp 142–147, Oct 2006

40. Janidarmian M, Roshan Fekr A, Samadi Bokharaei V (2011) Application-specific networks-on-chips design. IAENG Int J Comput Sci 38(1):16–25
41. Janidarmian M, Tinati M, Khademzadeh A, Ghavibazou M, Fekr AR (2010) Special issue on a fault tolerant network on chip architecture. AIP Conf Proc 1247:191–204
42. Wang LT, Stroud CE, Touba NA (2008) System-on-chip test architectures, nanometer design for testability, Mogran Kauffmann, Burlington
43. Dally WJ, Towles B (2003) Principles and practice of interconnection networks. Morgan Kaufmann, Burlington
44. Shamshiri S, Cheng KTT (2011) Modeling yield, cost, and quality of a spare-enhanced multi-core chip. IEEE Trans Comput

Chapter 27
Improved Anytime D* Algorithm

Weiya Yue, John Franco, Qiang Han and Weiwei Cao

Abstract In the dynamic domain, agents often operate in the terrain which is only incompletely known and can be dynamically updated on the fly. In this case, dynamic navigation algorithm, which is required to find out an optimal solution to its goal, has been an important component in planning. However, under the environments where time is more critical than optimality, a sub-optimal solution is required. Therefore the challenge for practical applications is to find a high sub-optimal solution in limited time. The dynamic algorithm Anytime D*(AD*) is currently the best anytime algorithm which aims to return a high sub-optimal solution with short corresponding time and control of sub-optimality. In this chapter, a new algorithm named Improved Anytime D*(IAD*) is introduced. By reducing the search space, experiment results show IAD* better outperforms Anytime D* in various random benchmarks.

Keywords Anytime algorithm · Artificial intelligence · D* lite · Incremental algorithm · Navigation algorithm · Planning

27.1 Introduction

With development of techniques, it is highly possible to develop autonomous vehicles, intelligent agents, etc. Because of this, navigation algorithm has been more and more important. In navigation algorithm, an agent is required to find out

W. Yue · J. Franco · Q. Han
Department of Computer Science, University of Cincinnati, Cincinnati, OH 45220, USA
e-mail: weiyayue@hotmail.com

W. Cao (✉)
Institute of Information Engneering, Chinese Academy of Science, Beijing 10009, China
e-mail: weiwei.cao@hotmail.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_27,

one optimal solution to its goal under changing environment. Many algorithms have been developed to solve this problem and gained big success [4, 5, 12, 13, 15, 16, 18]. But under some environments, where there is no sufficient resources for the agent to find out an optimal solution, a sub-optimal solution is acceptable. In this chapter, we will focus on the environment where time is more critical than optimality.

Time-limited search algorithms, called *anytime planning* algorithms, have been developed to fit in time critical environment. The basic idea is to find one solution as soon as possible, then to progressively replace a stored path with a better path when one is discovered during search until the time available for search expires [2, 3, 19]. Then the stored, probably sub-optimal path is the one that is used. In [2, 8, 11], it has been demonstrated that a so-called weighted A* algorithm variant which uses "inflated" heuristics (described below) can expand fewer vertices than the normal A* algorithm.

In A*, the vertices in *OPEN* are sorted by their values $f = g + h$. By assuming h is admissive, in A* algorithm if we use $f = g + \epsilon \cdot h$, then the returned path can be guaranteed to be ε sub-optimality, i.e. $g(v_g) \leq \varepsilon \cdot g^*(v_g)$ [1]. This strategy is called inflated heuristics, and the benefit is the control of ε sub-optimality. The A* algorithm using inflated heuristics is named weighted A* algorithm. In [2], one general method, named Anytime Weighted A*, to transform heuristic search algorithms to anytime algorithms is proposed. Anytime Weighted A* is an anytime planning algorithm which returns one sub-optimal solution as soon as possible, and then whenever allowed it continues to improve current solution until one optimal solution returned.

Anytime Weighted A* does initialization on most variants as A* algorithm, and as Weighted A*, the heuristic function h used is admissive. Different from normal A* algorithm, in Anytime Weighted A* p is used to record current returned path which may be improved later; *ERROR* is used to estimate how far away current solution is from optimal path; ε is the parameter used to inflate h; in one vertex $v \in OPEN$, the stored values are $\langle g(v), f'(v) \rangle$ instead of $\langle g(v), f(v) \rangle$, in which $f'(v) = g(v) + \varepsilon \cdot h(v)$. For example, in the priority queue *OPEN* of Anytime Weighted A*, the vertices are sorted by f' value instead of f value in normal A* algorithm. Although Anytime Weighted A* uses inflated heuristic value f' to sort vertices, normal f values are also recorded, which is used to prune searching space [3].

Under the changing environment, to search a path between a fixed pair of vertices within limited time, incremental anytime algorithm Anytime Repairing A*(ARA*) algorithm [10] was developed to mitigate this problem. ARA* runs weighted A* algorithm many times. Every time changes observed, ARA* needs to run weighted A* to find the new sub-optimal path. Most important to the performance of ARA* is that it reuses previously calculated information to avoid duplicating computation. This is done in accordance with ideas taken from [6, 7]. By observing that in weighted A* when h is admissive, if every vertex is allowed to be expanded only once, the returned path is still ε sub-optimal, every time

ARA* needs to recalculate, ARA* will only update a vertex at most one time. ARA* starts with a large value for the so-called *inflated parameter* ε and then reducing ε on each succeeding round until either $\varepsilon = 1$ or available time expires. ARA* performs similarly as Anytime Weighted A* algorithm and give the ability to control the sub-optimality ε.

D* Lite algorithm can be treated a dynamic version of lifelong A* algorithm [6, 7]. As D* algorithm [12, 13], D* lite searches backward from v_g to v_s. This is likely the critical point to the success of D* and its descendants because the g value of every node is exactly the path cost from that node to the goal v_g and can be used *after* the agent moves to its next position. The function *rhs* is defined by

$$rhs(v) = \begin{cases} \min_{v' \in succ(v)} g(v') + c(\langle v, v' \rangle) & v \neq v_g \\ 0 & \text{otherwise.} \end{cases}$$

The "more informed" *rhs* function assists in making better vertex updates during expansion. Call vertex v *locally consistent* if $rhs(v) = g(v)$, *locally overconsistent* if $rhs(v) < g(v)$, and *locally underconsistent* if $rhs(v) > g(v)$. In the latter two cases v is said to be *inconsistent*. A "best" path can be found if and only if, after expansion of v_s, all vertices on the path are locally consistent and can be computed by following the maximum-g-decrease-value vertices one by one from target v_g. If some changes that have been made since the last round cause a vertex v to become inconsistent then D* Lite will update $g(v)$ to make v locally consistent by setting $g(v) = rhs(v)$. Because D* Lite algorithm will only propagate inconsistent vertices to update partial vertices' (g, rhs) values instead of updating all vertices' (g, rhs) values, D* Lite can perform much better than other navigation algorithms.

Anytime D* [15] intends for dynamic navigation applications where optimality is not as critical as response time. It may be thought of as a descendant of both the Anytime Repairing A*(ARA*) and D* Lite algorithms. It may recalculate a best path more than once in a round with decreasing ε-suboptimality until $\varepsilon = 1$ or time has run out. Thus, Anytime D* will try to give a relatively good, available path quickly and, if time allows, will try to improve the path incrementally as is the case for Anytime A*.

In [15, 16], D* Lite algorithm has been improved by avoiding unnecessary calculations further, in which if the original optimal path is still available and can not be improved by changes observed, that path will be chose without computing. In [18], the ID* Lite algorithm uses a threshold number which is estimated to control the propagation of inconsistent vertices. Every time only changes may contribute a path whose weight is less or equal with the threshold number are propagated. The threshold number is increased until an optimal path found or all inconsistent vertices have been updated.

In this chapter, we will combine the techniques used in ID* Lite to improve Anytime D* algorithm and get an algorithm names Improved Anytime D* algorithm, and IAD* for abbreviation. In Sect. 27.2, pseudo code of IAD* is listed and described. Then in Sect. 27.3, IAD* and AD* are compared in various random benchmarks. At last, we conclude and discuss the next step of work. This work is a resivion of Chap. 26 [17].

27.2 Improved Anytime D* Algorithm

In this section at first we explain how IAD* algorithm works and then give its pseudo code. As Anytime D* algorithm, when changes observed, IAD* will try to update inconsistent vertices to get a new sub-optimal path. Every time, after recalculating, it is required to return a sub-optimal path whose weight is no bigger than $\varepsilon' \cdot g^*$ in which ε' is a preset parameter to control sub-optimality and g^* is the weight of current optimal path.

Every time, when recalculating needed, the sub-optimality parameter ε is reset to be ε' which is relatively big. By doing this, we expect to return a path as soon as possible [2, 8, 11]. In [14], it is recommended that in weighted A* ε can be set to be bigger than ε' and experiments show that the first path can be returned faster. This technique can be combined with any weighted A* algorithm easily. In order to speed up returning the first path, Anytime D* allow one vertex to be expanded at most one time which will be explained later. But it has been shown that in some benchmarks this may delay propagation of some critical vertices and slow down the algorithm [2]. In Sect. 27.3, we will run experiments to compare these results.

After the first path returned, as any other anytime algorithm, IAD* will try to improve current path until time runs out or an optimal path has been found. In Anytime D*, in order to do so, ε is decreased to look for a new path until $\varepsilon = 1$ which means the returned path is optimal. When ε is decreased, in Anytime D* all inconsistent are inserted into priority queue to be updated. As in ID* Lite [18], IAD* will not do any recalculating at all. Given a ε, at first, IAD* will try to find one consistent ε' sub-optimal path which is not affected by changes observed, and if such a path exists, it is returned immediately as the first path found. If no such a path found, IAD* will only choose part of overconsistent vertices whose propagation may lead to a ε sub-optimal path.

The pseudo code of IAD* is listed in Fig. 27.1. Function **Initialize()** defines and initializes ε, and the priority queues OPEN, CLOSED, and INCONS, and initialize values for g, *rhs* and *type* values of vertices. The initial value of ε_0 is relatively large in order to make sure some path is returned quickly. And vertex v_g and its *key* is inserted into priority queue OPEN.

In Function **key**(s), under-consistent vertices have their key-values updated as $g(s) + h(s)$ which is smaller than **key**(v_s). This processing can guarantee such kind of increasing changes can be propagated. Function **UpdateVertex**(s) updates one vertex in the same way as Anytime D* by using INCONS to store some of the inconsistent vertices and making sure that one vertex is expanded at most once in one execution of **ComputeOrImprovePath**(). After doing this, there is still returned solution generated satisfies ε-suboptimality [18].

Fig. 27.1 Main functions of IAD*

```
Improved Anytime D* Algorithm

Procedure Initialize()
OPEN = CLOSED = INCONS = catch=$\emptyset$;
for all $v \in V$, $rhs(v) = g(v) = \infty$; $\mathbf{type}(v) = -1$;
$rhs(v_g) = g(v_g) = \mathbf{type}(v_g) = 0$; $\varepsilon = \varepsilon_0$
OPEN.insert($[v_g,\ [h(v_g),\ 0]]$);

Procedure key($s$):
if $(g(s) > rhs(s))$
    return $[rhs(s) + \varepsilon \cdot h(s),\ rhs(s)]$;
else
    return $[g(s) + h(s),\ g(s)]$;

Procedure UpdateVertex($s$):
if $s$ has not been visited
    $g(s) = \infty$;
if $(s \neq v_g) rhs(s) = \min_{s' \in succ(s)}(c(\langle s, s' \rangle) + g(s'))$;
if ($s \in$ OPEN) OPEN.remove($s$);
if $(g(s) \neq rhs(s))$
    if ($s \in$ CLOSED)
        OPEN.insert($[s, \mathbf{key}(s)]$);$\mathbf{type}(v) = 0$;
    else
        insert $s$ into INCONS;

Procedure ComputeOrImprovePath():
while (OPEN.TopKey() $< \mathbf{key}(v_s)$ OR $rhs(v_s) \neq g(v_s)$)
    $s$ = OPEN.Top(), OPEN.remove($s$);
    if $(g(s) > rhs(s))$
        $g(s) = rhs(s)$;
        CLOSED.insert($s$);
        for all $s' \in pred(s)$ UpdateVertex($s'$);
    else
        $g(s) = \infty$;
        for all $s' \in pred(s) \cup \{s\}$ UpdateVertex($s'$);

Procedure MiniCompute( )
while (OPEN.TopKey() $< \mathbf{key}(v_c)$)
    $u$ = OPEN.Top(), OPEN.remove($u$);
    if $(g(u) > rhs(u))$
        $g(u) = rhs(u)$;
        CLOSED.insert($s$);
        for all $s' \in pred(s)$ UpdateVertex($s'$);
    else
        OPEN.Remove($u$);

Procedure GetAlternativePath($v_c$)
Vertex $r = v_c$;$C = \emptyset$
while $(r \neq v_g)$
    update $r$'s type value;
    if $(\mathbf{type}(r) > 0)$
        $r =$ one successor $y$ of $r$ with $rhs(y) + c(r,y) \leq rhs(r)$
          and $\mathbf{type}(y) \neq -3$ and $\mathbf{type}(y) \neq -2$;
    else if $(\mathbf{type}(r) == 0)$
        $\mathbf{type}(r)$ = -2;
        if $(r == v_c)$
            for every vertex $c \in C$ UpdateVertex($c$);
            return FALSE;
        $C = C \cup r's$ type value $-3$ children; $r = parent(r)$;
return TRUE.
```

Fig. 27.1 continued

```
Procedure GetBackVertex($v$)
01. if ($v \neq$ NULL and type($v$) $< 0$)
02.     if ($rhs(p) \neq g(p)$)
03.         return;
04.     type($v$) = 0;
05.     $v = parent(v)$;
06.     GetBackVertex($v$);

Procedure ProcessChanges()
01. Boolean $better$=FALSE, $recompute$ = FALSE, $t = rhs(v_c)$.
02. for every edge $e = \langle u, v \rangle$ where $c(e)$ has changed since the previous round:
03.     Update $rhs(u)$;
04.     if (type($u$) $= -3$) GetBackVertex($u$);
05.     if ($rhs(u) == g(u)$) type($u$) $= 0$;
06.     else
07.         if ($g(u) > rhs(u)$) and $\varepsilon * h(v_c, u) + rhs(u) < t$
08.             $better$ = TRUE, UpdateVertex($u$);
09.         else
10.             $catch$.add(u), type($u$) $= -3$;
11. if ($better ==$ TRUE) MiniCompute();
12. while (!GetAlternativePath($v_c$))
13.     $t_{old} = t$, ComputeShortestPath(), $t = rhs(v_c)$;
14.     if $t > t_{old}$
15.         $better$=FALSE;
16.         for every $u \in catch$ such that type($u$) $\neq 0$
17.             if ($\varepsilon * h(v_c, u) + rhs(u) < t$ and $g(u) > rhs(u)$)
18.                 $better$ = TRUE, UpdateVertex($u$);
19.                 $catch$.remove($u$).
20.         if ($better ==$ TRUE) MiniCompute().

Procedure Main():
01. Initialize();
02. ComputeOrImprovePath();GetAlternativePath($v_c$);
03. publish current $\varepsilon$-suboptimal solution;
04. repeat the following:
05.     for all directed edges $\langle u, v \rangle$ with changed edge costs
06.         Update the edge cost $c(\langle u, v \rangle)$;
07.         UpdateVertex($u$);
08.     if significant edge cost changes were observed
09.         increase $\varepsilon$ or replan from scratch;
10.     else if ($\varepsilon > 1$)
11.         decrease $\varepsilon$;
12.     CLOSED = $\emptyset$;
13.         ProcessChanges();
14.     publish current $\varepsilon$-suboptimal solution;
15.     if ($\varepsilon == 1$)
16.         wait for changes in edge costs;

Procedure MoveAgent():
01. while ($v_s \neq v_g$)
02.     wait until a plan is available;
03.     Set type($v_c$) $= 0$;
04.     $v_c = u$ where $u$ is a successor of $v_c$ and type($u$) $> 0$;
05.     Move the agent to $v_c$;
```

Functions **ComputeorImprovePath**(), **MiniCompute**() and **GetBackVertex(v)** are the same as in ID* Lite. Function **GetAlternativePath**(v_c) returns TRUE if and only if there is a path from v_c to v_g and, if it returns TRUE, it has changed type values on vertices so that a least cost path from v_c to v_g can be traversed by visiting neighboring vertices of lowest positive type until v_g is reached. Different from ID* Lite, at line 05, instead of choosing a child of r, one successor y of r with $rhs(y) + c(r, y) \leq rhs(r)$ is chose. The reason is that here we only need a suboptimal path.

Different from ID* Lite, here the returned path may be not optimal, but is guaranteed to be ε' suboptimal. It worths to notice that if a path returned, and on which there are overconsistent vertices, then the path returned is better than ε'

suboptimal. If no path can be returned by function **GetAlternativePath**(v_c), every vertex $c \in C$ is updated by function **UpdateVertex**. Observe that c is underconsistent and that this is the only place in the code where underconsistent vertices are placed in OPEN as in ID* Lite [18]. This is because only increased changes will cause underconsistent vertices, and increased changes are only inserted here. Decreased changes have been inserted before **GetAlternativePath**(v_c) was called.

ProcessChanges acts similarly as ID* Lite. But differently, at line 07 and 17, to test whether a overconsistent vertex should be put in OPEN to propagate, $\varepsilon * h(v_c, u) + rhs(u) < t$ is used instead of $h(v_c, u) + rhs(u) < t$. Functions **Main**() and **MoveAgent**() are the same as Anytime D*.

We end this Section with the Theorem of correctness of IAD*.

Theorem 1 In Improved Anytime D* algorithm, the returned path between v_c and v_g has its cost no larger than $\varepsilon * g'(v_c)$ in which $g'(v_c)$ is the cost of optimal path between v_c and v_g.

Proof This Theorem follows the correctness of ID* Lite algorithm and Anytime D* algorithm.

27.3 Experiments and Analysis

In this section, the performance of IAD* is compared experimentally with Anytime D* on random grid world terrains. In each experiment the terrain is a square, 8-direction grid world of $size^2$ vertices. Special vertices v_s and v_g are chosen randomly from the terrain. Initially $percent\% * size^2$ of the vertices are selected randomly and blocked, *percent* being a controlled parameter. The parameter *sensor-radius* is used to set the maximum distance to a vertex that is observable from the current agent position. Before navigation, the traveling agent has a old map, in which an obstacle may be wrongly considered to be blank with a fifty percent possibility.

To compare the anytime dynamic navigation algorithms, we control the sub-optimality by setting parameter ε. I.e. solutions returned by the algorithm is ε sub-optimal. One criterion of anytime algorithm is the time of returning the first sub-optimal path. In this condition, the less operations used, the better is the algorithm. We will run two kinds of random benchmarks to compare IAD* algorithm and Anytime D* algorithm with constant sub-optimality parameter ε. The first set of results are on random rock-and-garden benchmarks. That is, a blockage is set initially and will remain for the entire experiment. The second set of results are on a collection of benchmarks that model agent navigation through changing terrain, named as Parking-lot benchmarks. I.e., a blockage may move to its neighborhood randomly during the navigation.

Form Figs. 27.2, 27.3, 27.4 and 27.5, AD* and IAD* are compared on rock-and-garden benchmarks. From Figs. 27.2 to 27.5, the results of $percent = 10$ and $sensor - radius = 0.1 * size$ are presented for $size = 300$ and $size = 500$

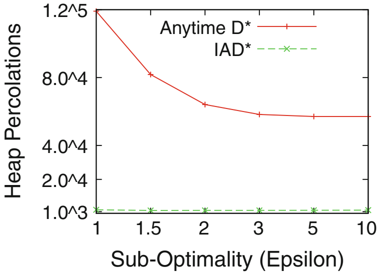


Fig. 27.2 size = 300, percent = 10, sensor − radius = 30(rock-and-garden)

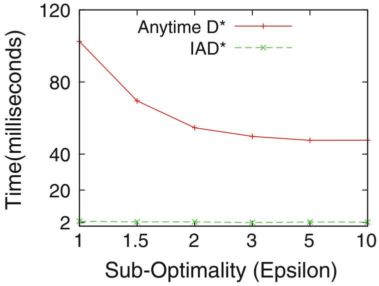


Fig. 27.3 size = 500, percent = 10, sensor − radius = 50(rock-and-garden)

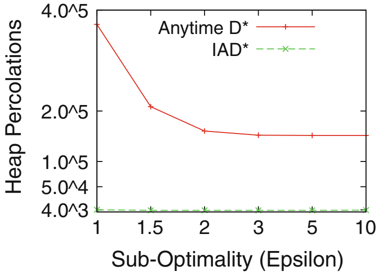


Fig. 27.4 size = 500, percent = 10, sensor − radius = 50(rock-and-garden)

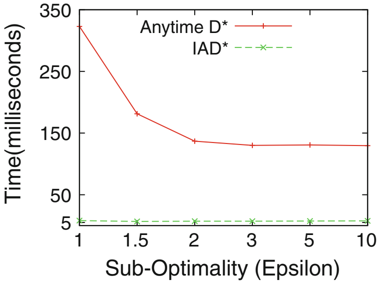


Fig. 27.5 size = 300, percent = 10, sensor − radius = 30(rock-and-garden)

respectively. These figures show the number of heap operations and also the time consumed as a function of sub-optimality EPSILON(ε). From the figures, we can see that IAD* gains more than one order of speeding up than AD* in some cases. About the consumed time, IAD* needs extra time to calculate the alternative path besides the time of applying heap operations which is different from AD*. From the figures we can see that in IAD* the time to calculate alternatives cost only takes a very small percentage of the whole time, which means it does not affect the speed of IAD* much. It also shows that heap operations consist the main time complexity in navigation algorithms. Compared with the speeding up of ID* Lite for D* Lite [18], the speed up performance of IAD* for AD* is much better. There are two main reasons, the first one is that there are more alternatives because of sub-optimality. In order to find an alternative of path P_1, in ID* Lite, only paths with the same cost as P_1 can be used as alternatives and there are no paths with cost smaller than P_1. But in IAD*, all paths with cost $\leq P_1$ can be used as alternatives. The second one is that AD* needs to reorder its priority queue OPEN every time of recalculating. If IAD* can avoid the recalculating, then the reordering of priority queue can be skipped. The method used in [12] to avoid reordering priority queue can also be used in AD* by modifications introduced in [18]. Unfortunately, the heuristic used in AD* is not consistent, hence that method will cause a lot of reinsertion of key values of vertices because of updating. In [18], the authors choose to reorder the priority queue when recalculating and consider this operation is bearable for time.

The second set of benchmarks, Parking-lot benchmarks, is intended to model agent navigating in the presence of terrain changes. Compared with rock-and-garden benchmarks, we are more interested in parking-lot benchmarks. The reason is the later can simulate the practical environment better. Hence we will give more comparisons between IAD* and AD* on this kind of benchmarks. Terrain changes are commonly encountered by autonomous vehicles of all kinds and may represent the movement of other vehicles and structures in the agent's environment. A number of *tokens* equal to a given fixed percentage of vertices are initially created and distributed over vertices in the grid, at most one token covering any vertex. As an agent moves from vertex to vertex through the grid tokens move vertex to vertex as well. Tokens are never destroyed or removed from the grid and the rules for moving tokens do not change: on each round a token on vertex v moves to a vertex adjacent to v with probability 0.5 and the particular vertex it moves to is determined randomly and uniformly from the set of all adjacent vertices that do not contain a token when the token is moved. Tokens are moved sequentially so there is never more than one token on a vertex. Any vertex covered by a token at any point in the simulation is considered blocked at that point which means all edge costs into the vertex equal ∞. A vertex with no token is unblocked and edge costs into it are not ∞.

From Figs. 27.6, 27.7, 27.8 and 27.9, IAD* and AD* are compared similarly as from Figs. 27.2 to 27.5. We can see that IAD* can also get one order of speeding up, but not as good as in rock-and-garden benchmarks. The reason is that in parking-lot benchmarks, more recalculating are needed by IAD*. This is similar as ID* Lite compared with D* Lite [18]. As we have seen, the time has similar plot as heap percolation, hence below we only show the heap percolation plot because of the limited space.

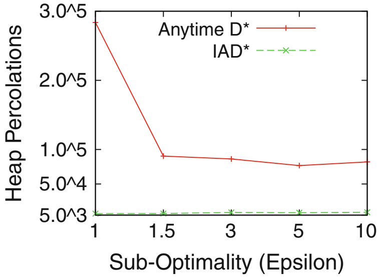


Fig. 27.6 size = 300, percent = 10, sensor − radius = 30(Parking-lot)

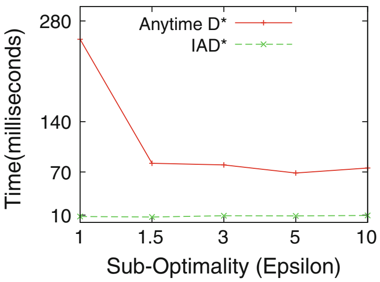


Fig. 27.7 size = 300, percent = 10, sensor − radius = 30(Parking-lot)

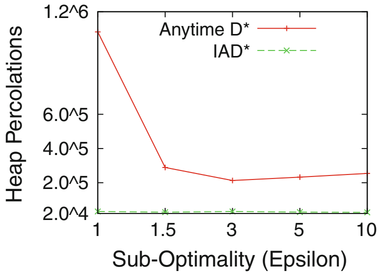


Fig. 27.8 size = 500, percent = 10, sensor − radius = 50(Parking-lot)

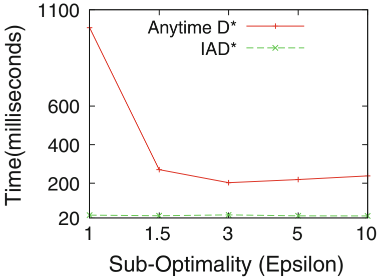


Fig. 27.9 size = 500, percent = 10, sensor − radius = 50(Parking-lot)

In Figs. 27.10 and 27.11, AD* and IAD* are compared similarly as in Figs. 27.6 and 27.8, but with *percent* = 20. This is to show whether IAD* can perform well in dramatically changing environments. We can see IAD* can still achieve up to one order times of speeding up than AD*. Also, from such figures we

Fig. 27.10 size = 300, percent = 20, sensor − radius = 30(Parking-lot)

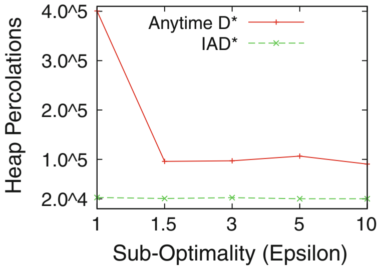


Fig. 27.11 size = 500, percent = 20, sensor − radius = 50(Parking-lot)

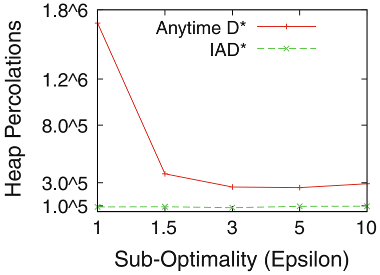
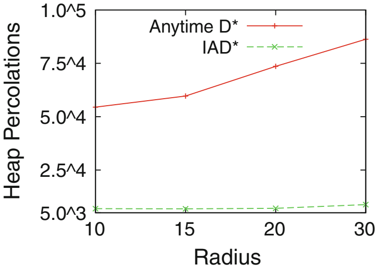


Fig. 27.12 size = 300, percent = 10, ε = 3(Parking-lot)



can see that the plot of AD* decreases faster than IAD*. Hence we can say that with other conditions made certain, the smaller the sub-optimality required, the better IAD* performs.

In Fig. 27.12, results are compared with different sensor-radius given $size = 300, percent = 3$ and $\varepsilon = 3$. We can see the heap operations increases with $sensor - raduis$. The reason is that in parking-lot benchmarks the blockages are keeping moving, hence a larger sensor-radius means more changes observed which may cause more updating.

In Fig. 27.13, results are compared with different *percent* given $size = 300, sensor - radius = 30$ and $\varepsilon = 3$. We can see that IAD* has heap operations increased faster than AD* when percent is increased. When percent is increased, in order to find an alternative, more recalculations are needed in IAD*, which affects its performance. Hence, in parking-lot kind of environments, if the changing of terrain is relatively light, IAD* can have a better performance than in heavily

Fig. 27.13 size = 300, sensor − radius = 30, ε = 3(Parking-lot)

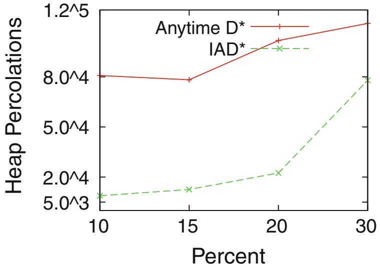
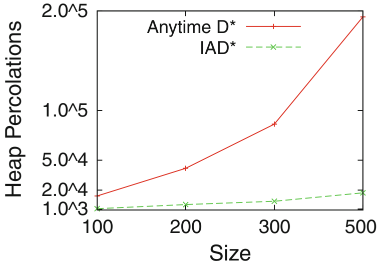


Fig. 27.14 percent = 10, sensor − radius = 30, ε = 3(Parking-lot)



changing environments. From the figure, we can also see that, even with *percent* = 30, IAD* can still get about two times speeding up than AD*.

At last in Fig. 27.14, results are compared with different *size* given *percent* = 10, *sensor* − *radius* = 30 and ε = 3. The results show that IAD* is very scalable in parking-lot benchmarks.

From the results above, we can see IAD* returns the first qualified sub-optimal path in a shorter time than AD*. Also when other conditions unchanged and in rock-and-garden benchmarks ε from 1 to 1.5 IAD* can save about 25 percent of calculations; but in parking-lot benchmarks, the calculations of IAD* varies a little. So IAD* has the ability of returning a high sub-optimal path especially in parking-lot style of environments, and when something emergency happens, for example a lot of changes observed, the first sub-optimal path can be returned fast; after that, IAD* can continue to improve current path until time runs out. I.e., a desired sub-optimal path can be guaranteed to be found in a short time with a high possibility, which also means more time can be used to improve the firstly returned path. Hence, we can conclude that IAD* can return the first sub-optimal path faster than AD* in various random benchmarks, from which IAD* gains a better potentiality to return high sub-optimal path within limited time.

27.4 Conclusion and Next Step of Work

In this chapter, we propose a new dynamic anytime algorithm IAD*. IAD* improves AD* following the similar strategy as that ID* Lite improves D* Lite.

That is, IAD* will try to find an alternative of original path instead of recalculating immediately as in AD*. Moreover, if an alternative is not available, in order to avoid a full recalculation IAD* will try to propagate changes part by part with the help of a threshold until a new sub-optimal path found. Experimental results show that IAD* can achieve up to one order of speeding up in various random benchmarks. There is still much work can be done in the next step. For example, We will consider when there is an upper bound of heap operations allowed if there are limited resource and how to make the algorithm to achieve a better sub-optimality; As discussed in this chapter IAD* has a better potential to support high sub-optimal path, so we can compare this aspect of AD* and IAD* by experiment; Anytime Weighted A*(AWA*) [2] is demonstrated performing better in some benchmarks to achieve a higher sub-optimality than other algorithms, so combining IAD* with AWA* to get a new faster algorithm is also an interesting work we will do.

Acknowledgments This work was supported by University Research Council's (URC), 2012 Summer Graduate Student Research Fellowships funded by University of Cincinnati.

References

1. Davis HW, Bramanti-Gregor A, Wang J (1988) The advantages of using depth and breadth components in heuristic search. Methodol Intell Syst 3:19–28
2. Hansen EA, Zhou R (2007) Anytime heuristic search. J Artif Intell Res 28:267–297
3. Harris L (1974) The heuristic search under conditions of error. Artif Intell 5(3):217–234
4. Koenig S, Likhachev M (2002) D*lite. In: Eighteenth national conference on, artificial intelligence, pp 476–483
5. S. Koenig, Likhachev M (2002) Improved fast replanning for robot navigation in unknown, Terrain, pp 968–975
6. Koenig S, Likhachev M (2002) Incremental a*. Advances in neural information processing systems, pp 1539–1546
7. Koenig S, Likhachev M, Furcy D (2004) Lifelong planning a*. Artif Intell J 155(1–2):93–146
8. Korf R (1993) Linear-space best-first search. Artif Intell 62(1):41–78
9. Likhachev M, Ferguson D, Gordon G, Stentz A, Thrun S (2005) Anytime dynamic a*: an anytime, replanning algorithm. In: Proceedings of the international conference on automated planning and scheduling
10. Likhachev M, Gordon G, Thrun S (2003) Ara*: anytime a* with provable bounds on sub-optimality. Adv Neural Inf Process Syst
11. Pohl I (1970) Heuristic search viewed as path finding in a graph. Artif Intell 1(3):193–204
12. Stentz A (1995) The focussed d* algorithm for real-time replanning. In: Proceedings of the international joint conference on, artificial intelligence, pp 1652–1659
13. Stentz A (1997) Optimal and efficient path planning for partially-known environments, vol 388. The Kluwer international series in engineering and computer science, pp 203–220
14. Thayer JT, Ruml W (2008) Faster than weighted a*: an optimal approach to bounded suboptimal search. In: Proceedings of the international conference on automated planning and scheduling
15. Yue W, Franco J (2009) Avoiding unnecessary calculations in robot navigation. In: Proceedings of world congress on engineering and computer science, pp 718–723

16. Yue W, Franco J (2010) A new way to reduce computing in navigation algorithm. J Eng Lett 18(4): EL _18 _4 _03
17. Yue W, Franco J, Cao W, Han Q (2012) A new anytime dynamic navigation algorithm. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, vol 1, pp 17–22, San Francisco, USA, 24–26 Oct 2012
18. Yue W, Franco J, Cao W, Yue H (2011) Id* lite: improved d* lite algorithm. In: Proceedings of 26th symposium on applied, computing, pp 1364–1369
19. Zhou R, Hansen E (2002) Multiple sequence alignment using anytime a*. In: Proceedings of conference on articial, intelligence, pp 975–976

Chapter 28
Scheduling of Real-Time Networks with a Column Generation Approach

Ernst Althaus, Sebastian Hoffmann, Joschka Kupilas and Eike Thaden

Abstract We present an algorithm based on column generation for the real-time scheduling problem of allocating periodic tasks to electronic control units in multiple subsystems connected by a global bus. The allocation has to ensure that tasks can be scheduled, and messages between tasks in different subsystems can be transmitted over the global bus and meet their deadlines. Also tasks and messages occurring in a task chain must be scheduled in a way such that the sequence of execution meets their end-to-end deadline. We show that our approach computes the optimal allocation in our model and due to the column generation approach early provides lower bounds on the optimal value.

Keywords Column-generation · Deadline-monotonic-scheduling · End-to-end-deadline · Rate-monotonic-scheduling · Real-time-network · Scheduling · Task-chain

E. Althaus · J. Kupilas
Max-Planck-Institut für Informatik, Campus E14, 66123 Saarbrücken, Germany
e-mail: jkupilas@mpi-inf.mpg.de

E. Althaus · S. Hoffmann (✉)
Institut für Informatik, Johannes Gutenberg-Universität, Staudingerweg 9,
55128 Mainz, Germany
e-mail: sebastian.hoffmann@uni-mainz.de

E. Althaus
e-mail: ernst.althaus@mpi-inf.mpg.de

E. Thaden
Department für Informatik, Carl von Ossietzky-Universität, 26111 Oldenburg, Germany
e-mail: eike.thaden@informatik.uni-oldenburg.de

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_28,

28.1 Introduction

In this chapter we look at the scheduling problem arising in the field of manufacturing embedded systems. We are given a hardware architecture with a global bus system, e.g. a FlexRay bus, connecting several processing subsystems $S = \{s_1, s_2, \ldots, s_k\}$. Each electronic control unit (ECU) in a subsystem can be of a specific ECU type $\{et_1, et_2, \ldots, et_\ell\}$, see Fig. 28.1.

Besides a set $T = \{t_1, t_2, \ldots, t_n\}$ of tasks specified by their worst-case execution time (WCET), deadline, period, and memory consumption, where the WCET and memory consumption depend on the ECU type, we are given a set $M = \{m_1, m_2, \ldots, m_v\}$ of messages specified by a single source task, a set of destination tasks, a transmission time, and a deadline, see Fig. 28.4. The scheduling problem arises in allocating every task to a specific ECU in a subsystem such that every message regarding this task can be transmitted over the global bus and both tasks and messages meet their deadlines in every periodic cycle.

We assume that tasks and messages arrive with a fixed rate given as their period, which is greater than or equal to their deadline—the period of a message is determined by the source task—and that we are given deadline monotonic priorities, which was proven to be optimal in our setting [13]. Each task can only be assigned to one ECU and each ECU can execute exactly one task at a time by using preemptive fixed-priority scheduling. Furthermore, the WCET and memory consumption of a task depend on the chosen ECU type of the ECU the task is allocated to. We call a schedule feasible if every task is entirely executed before its deadline.

We present two sufficient models to determine the feasibility of a schedule: first, the computation of the first idle time per ECU by Lehoczky et al. [12] and second, by the computation of the worst-case response time (WCRT) by the well-known fix-point equation by Joseph and Pandya [11].

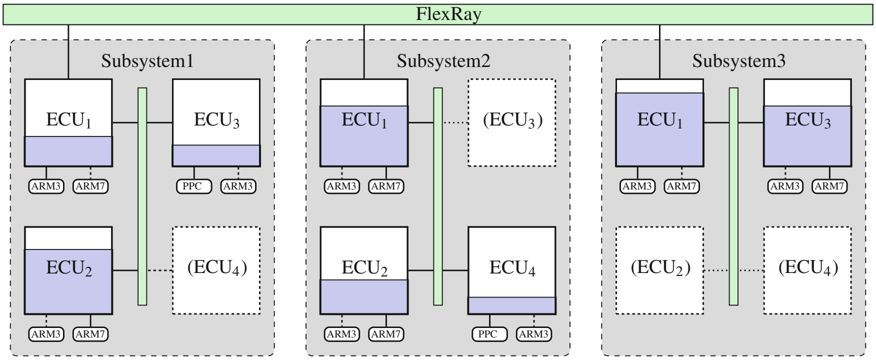


Fig. 28.1 A hardware architecture with three subsystems connected by a global FlexRay bus via the specific gateway ECUs in each subsystem. Upgrading the ECUs increases the cost that has to be minimized

There are three common bus systems that can be considered for the communication: a token area network (TAN) bus, a controller area network (CAN) bus or a FlexRay bus. We assume that the message communication in every ECU and subsystem is guaranteed and specific gateway ECUs are not necessary. We model a global FlexRay bus so that a message has to be sent over the FlexRay if at least one of the destination tasks is allocated to a different subsystem than the source task.

The problem at hand is of major importance in several industrial sectors, e.g. in aerospace, automotive, and automation industries, as it can save a lot of costs by optimally allocating tasks that have to be definitely executed.

In this chapter we formulate the problem as an integer linear program (ILP) and solve it by a column generation approach within a branch and bound framework. We extend the approach of Althaus et al. [2] by integrating the subsystems to the formulation, accelerating the computation if no response times are necessary, handling task chains with their end-to-end deadlines and modeling a global FlexRay bus.

28.2 Previous Work

The complete design flow from specification models to their distributed execution on hierarchical architectures with different approaches of pre-allocating tasks is described in Bücker et al. [5] and Clark et al. [7] whose repetitive two-tier approach first heuristically distributes the tasks to subsystems, and second, exactly solves the scheduling problem in every subsystem.

Eisenbrand et al. [9] formulated the problem of scheduling pre-allocated tasks in a subsystem as an integer linear program (ILP) which is solved by a standard ILP solver. Althaus et al. [2] improved upon their work by performing a Dantzig–Wolfe decomposition of the ILP formulation by introducing a column generation approach and obtained better running times on large instances.

28.3 A Column Generation Approach for the Scheduling Model

In this section we describe our column generation approach for solving the previously defined scheduling problem by presenting an ILP formulation following Althaus et al. [2].

Given a set $S = \{s_1, s_2, \ldots, s_k\}$ of subsystems, we call a subset $p \subseteq T$ of tasks a *task pattern* for subsystem $s \in S$ if all tasks in p can be scheduled in the subsystem s and we denote by $P^s \subseteq \mathscr{P}(T)$ the set of all task patterns for s, where $\mathscr{P}$ denotes the power set. Associated with each $p \in P^s$ in the ILP formulation we introduce a binary variable $X_{s,p}$ indicating whether the task pattern p is used for the schedule in subsystem s or not.

The objective function (28.1) is to minimize the sum of costs $cost(s,p)$ resulting from the choice of ECU types in each subsystem s needed to execute the allocated task pattern p. We impose two requirements for the task patterns: first, exactly one task pattern is used in a subsystem (28.2), and second, each task has to be assigned to exactly one subsystem (28.3).

Similarly to the set of task patterns P^s we define the set of *message patterns* $Q \subseteq \mathscr{P}(M)$ where $q \in Q$ if all messages in q can be transmitted over the global bus with respect to some bus protocol. We likewise introduce a binary decision variable Z_q for every $q \in Q$ with the requirement that exactly one message pattern must be chosen (28.6).

Furthermore, we introduce binary variables Y_m for every message $m \in M$ indicating whether the message m is sent over the global bus. Therefore, Y_m can only be 0 if all destination tasks $m_B \subseteq T$ of message m are in the same subsystem as the source task $m_a \in T$ (28.4). The constraint is not formulated as equality as there can exist broadcasting messages that need to be on the bus anyway. By Eq. (28.5) we ensure that the chosen message pattern contains all messages that have to be transmitted over the global bus.

Thus, our scheduling problem can be formulated as the following ILP:

$$\min \sum_{s\in S} \sum_{p\in P^s} cost(s,p) \cdot X_{s,p} \tag{28.1}$$

$$s.t. \sum_{p\in P^s} X_{s,p} = 1 \qquad \forall\, s \in S \tag{28.2}$$

$$\sum_{s\in S} \sum_{p\in P^s \,:\, t\in p} X_{s,p} = 1 \qquad \forall\, t \in T \tag{28.3}$$

$$-Y_m + \sum_{s\in S} \sum_{\substack{p \in P^s: \\ m_a \in p \ \wedge \ m_B \setminus p \neq \emptyset}} X_{s,p} \leq 0 \quad \forall\, m \in M \tag{28.4}$$

$$-Y_m + \sum_{q\in Q \,:\, m\in q} Z_q \geq 0 \qquad \forall\, m \in M \tag{28.5}$$

$$\sum_{q\in Q} Z_q = 1 \tag{28.6}$$

$$X_{s,p} \in \{0,1\} \qquad \forall\, s \in S, p \in P^s \tag{28.7}$$

$$Z_q \in \{0,1\} \qquad \forall\, q \in Q \tag{28.8}$$

$$Y_m \in \{0,1\} \qquad \forall\, m \in M \tag{28.9}$$

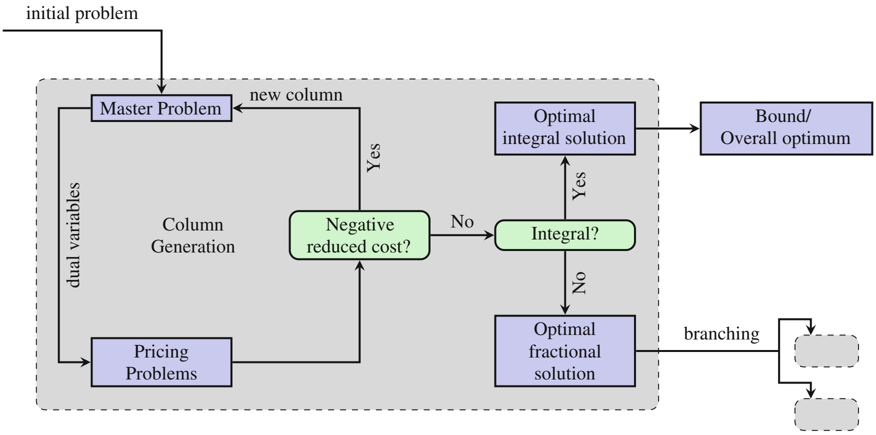


Fig. 28.2 Our column generation approach inside a branch and bound framework. Computed optimal integral solutions can be used as bounds in the master problems to prune the search tree and reduce the running time

Applying a branch and bound approach, we get bounds from solving the LP relaxations of the ILPs occurring in the branching process. Since the pattern variables arise in exponential number, we use a column generation approach to solve the LP relaxations, see Fig. 28.2.

The remainder of Sect. 28.3 is organized as follows. In Sect. 28.3.1 we expand on the LP relaxation. In Sect. 28.3.2 we explain the branch and bound algorithm. In Sect. 28.3.3 we present the models to scheduling periodic real-time tasks. In Sect. 28.3.4 we show how to handle task chains with end-to-end deadlines. In Sect. 28.3.5 we expand on the FlexRay bus for the global communication.

28.3.1 Solving the Relaxation

We relax the integrality constraints (28.7) of the task patterns $X_{s,p} \in \mathbb{R}$ and add the lower bounds $X_{s,p} \geq 0$ since the upper bounds $X_{s,p} \leq 1$ are implied by the constraints (28.2).

To state the so-called *master problem*, we make the problem artificially feasible by introducing a *super subsystem* $\tilde{s} \notin S$ in which tasks are only executable altogether, i.e. $P^{\tilde{s}} = \{T\}$. The feasibility of the original problem can be decided by inspecting the objective value or the new pattern variable $X_{\tilde{s},T}$, because we assign a cost higher than every optimal solution to it. If there are broadcasting messages we also have to generate an artificial message pattern that satisfies (28.6), but forces $X_{\tilde{s},T}$ to be 1 if the former is selected, if not, the empty message pattern $Z_{\emptyset} \in Q$, where no message is send over the global bus, will satisfy (28.6).

In the latter case the constructed master problem is feasible by default with

$$Z_\emptyset = X_{\tilde{s},T} = X_{s,\emptyset} = 1 \quad \text{and} \quad Y_m = 0$$

for all $s \in S$ and $m \in M$ with the objective value $cost(\tilde{s}, T)$.

The master problem is solved with a state-of-the-art simplex method and the so-called *pricing problems* are solved to check if non-basic variables $X_{s,p}$ or Z_q with negative reduced cost can be produced. In this case, the solution of the master problem is not optimal for the original problem—apart from degeneracies—and we have to add the variable to the master problem by generating a new column. This step is repeated until there are no more non-basic variables with negative reduced cost and the last solution to the master problem is also optimal for the original problem, see Fig. 28.2.

The structure of the problem allows us to partition the search for new variables into several independent pricing problems, namely into the *task pricing problems*, the search for a variable $X_{s,p}$ with the most negative reduced cost for a given subsystem s (28.10), and the *message pricing problem*, the search for a variable Z_q with the most negative reduced cost (28.15).

First, we generate the task pricing problem for a given subsystem s. By writing d_s for $s \in S$ for the dual variables corresponding to the subsystem constraints (28.2), d_t for $t \in T$ corresponding to the task constraints (28.3), and d_m for $m \in M$ corresponding to the message constraints (28.4) the pricing problem for a given $s \in S$ reads as follows:

$$\min \; cost(s,p) - d_s - \sum_{t \in T} d_t \cdot p_t - \sum_{m \in M} d_m \cdot y_m \tag{28.10}$$

$$s.t. \, y_m - p_{m_a} + p_b \geq 0 \qquad \forall\, m \in M, b \in m_B \tag{28.11}$$

$$p = [p_t]_{t \in T} \in P^s \tag{28.12}$$

$$p_t \in \{0,1\} \qquad \forall\, t \in T \tag{28.13}$$

$$y_m \in \{0,1\} \qquad \forall\, m \in M \tag{28.14}$$

with binary variables p_t indicating whether task t is represented in the task pattern p, and y_m indicating whether message m has to be sent over the global bus. Notice that the constraints (28.11) for a given message $m = (m_a, m_B) \in M$ force y_m to be 1, if the source task m_a is assigned to this subsystem and one of the destination tasks $b \in m_B$ is not.

Similarly, we generate the pricing problem for the message pattern by writing d'_m for the dual variable corresponding to the message constraints (28.5), and d_Q corresponding to the message pattern constraint (28.6) to obtain

$$\min \; -\sum_{m \in M} d'_m \cdot q_m - d_Q \tag{28.15}$$

$$\text{s.t.} \;\; q = [q_m]_{m \in M} \in Q \tag{28.16}$$

$$q_m \in \{0, 1\} \quad \forall\, m \in M \tag{28.17}$$

with binary variables q_m indicating whether message m is transmitted over the global bus.

The constraint (28.12) has to express the schedulability of the tasks in a subsystem, see Sect. 28.3.3, as well as the constraint (28.16) has to ensure the global message transmission due to the different bus types, see Sect. 28.3.5.

28.3.2 Branch and Bound

By solving the LP relaxation we obtain a lower bound on the optimal objective value of the ILP. If the cost of this solution is not smaller than the best integral solution found so far, we can stop. Otherwise, we branch by identifying a fractional assigned task to more than one subsystem and generate two branches: first, we predeploy this task to the specific subsystem, and second, we forbid this task to run in this subsystem. The process sequence of the branches is given by a priority queue in the order of the smallest lower bound first.

We achieve a big improvement in the runtime if we reuse columns as starting solutions that do not contradict the branching rule. To interrupt the process of solving pricing problems early we use the Lagrangian bound of Althaus et al. [2] in case that the current objective value cannot be further improved. Additionally, we trade on the integrality of the objective function by rounding up the lower bounds.

28.3.3 Scheduling Periodic Tasks

We are given tasks $t \in T$ with WCET c_t, deadline d_t, and period π_t satisfying

$$c_t \leq d_t \leq \pi_t$$

and we search for a fixed allocation of priorities to the tasks, a so called fixed priority scheduling (FPS) policy.

Eisenbrand et al. [9] state a necessary and sufficient schedulability test by computing the exact WCRTs of the tasks with an ILP formulation and bound them by the specific deadlines due to the well-known recursive equation from Joseph and Pandya [11].

In the given case of non-implicit deadlines, i.e. $d_t \leq \pi_t$, the deadline monotonic scheduling (DMS) policy, i.e. the priority is inversely proportional to its deadline, is proven to be optimal under all FPS policies, see Burns and Wellings [6].

In the special case of implicit deadlines, i.e. $d_t = \pi_t$, the DMS policy equals the rate monotonic scheduling (RMS) policy which is therefore also optimal under all FPS policies.

A sufficient but not necessary schedulability test is provided by Audsley [3] which becomes sufficient and necessary if and only if the considered WCRTs are exact. The workload analysis by Lehoczky et al. [12] presents a pseudo-polynomial, necessary and sufficient scheduling test without computing the WCRTs which can also be stated as linear constraints.

The fact that the WCET and memory consumption depend on the chosen ECU type of the ECU the task is assigned to leads to high complexity for the occurring ILPs. To present the concept of Lehoczky et al. we ignore the latter dependencies and assume that there is only one ECU of a fixed ECU type.

For a given task $t \in T$ we compute the *cumulative workload demands* on the ECU over an interval $[0, \delta]$, where 0 is a critical instant for the set $\{t' \mid t' \in hep(t)\}$, and where $hep(t)$ denotes all tasks with priority higher than or equal to t, i.e. $d_{t'} \leq d_t$, by

$$W_t(\delta) = \sum_{t' \in hep(t)} c_{t'} \cdot \left\lceil \frac{\delta}{\pi_{t'}} \right\rceil. \tag{28.18}$$

Lehoczky et al. [12] show that the entire task set can be scheduled if

$$\max_{t \in T} \min_{\delta \in \Delta_t} \sum_{t' \in hep(t)} \frac{c_{t'}}{\delta} \cdot \left\lceil \frac{\delta}{\pi_{t'}} \right\rceil \leq 1 \tag{28.19}$$

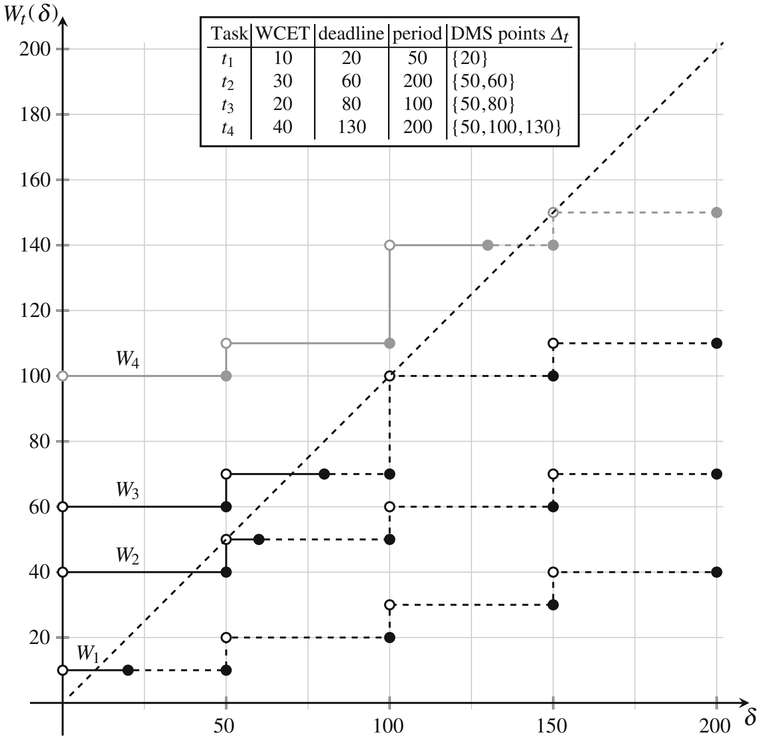
with the so-called *deadline monotonic scheduling points*

$$\Delta_t = \{d_t\} \cup \{k \cdot \pi_{t'} \leq d_t \mid t' \in hep(t), k \in \mathbb{N}\} \tag{28.20}$$

for task t, see Fig. 28.3. To replace (28.12) in the task pricing problem we express (28.19) by the following constraints (28.21) and (28.22):

$$-M_t \cdot \alpha_{\delta,t} + \sum_{t' \in hep(t)} \frac{c_{t'}}{\delta} \cdot \left\lceil \frac{\delta}{\pi_{t'}} \right\rceil \cdot p_{t'} \leq 1 \tag{28.21}$$

with binary variables $\alpha_{\delta,t} \in \{0, 1\}$ for all $t \in T$ and $\delta \in \Delta_t$ with M_t sufficiently big.



Task	WCET	deadline	period	DMS points Δ_t
t_1	10	20	50	{20}
t_2	30	60	200	{50,60}
t_3	20	80	100	{50,80}
t_4	40	130	200	{50,100,130}

Fig. 28.3 Cumulative workload demand function $W_t(\delta)$ for some tasks in the interval $[0, \delta]$. It is obvious that (28.19) has to be checked only in the DMS points Δ_t for a schedulability analysis to see that task t_4 cannot be added if the other three tasks are assigned

The constraints

$$\sum_{\delta \in \Delta_t} \alpha_{\delta,t} \leq |\Delta_t| - 1 \quad \forall\, t \in T \tag{28.22}$$

ensure that there has to be at least one point $\delta \in \Delta_t$ where it is not necessary to subtract M_t from the demand to satisfy (28.19), resp. (28.21).

28.3.4 End-to-end Deadlines for Task Chains

A task chain $c \subseteq T$ is a set of related tasks with their corresponding messages that has to be executed within a time restriction of an end-to-end deadline. We restrict to the case of simple linear task chains, i.e. $|m_B \cap c| \leq 1$ for every message m with $m_a \in c$, which can always be obtained from more complex ones, see Fig. 28.4. In addition, we assume that every message between tasks in a task chain has to be considered in the end-to-end deadline.

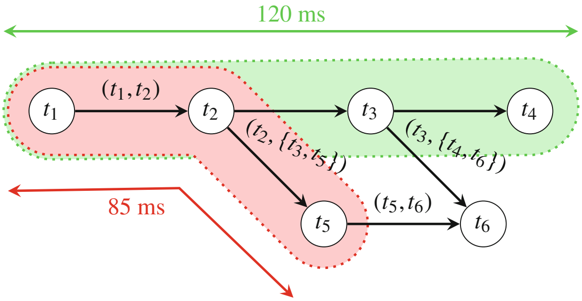


Fig. 28.4 A small real-time network with two simple linear task chains and their end-to-end deadlines. One sequence starting from source task t_1 via t_2, t_3 and t_4 must be finished within the end-to-end deadline of 120ms, the other sequence from t_1 to t_5 via t_2 must be scheduled within 85ms. The message $(t_2, \{t_3, t_5\})$ will be splitted in the computation

Since tasks of a chain can be distributed over several subsystems, the upcoming messages have to be transmitted over the global bus and therefore the computation of the worst-case end-to-end latency has to take into account the WCRT of all involved tasks and global messages in the chain as well as the periods of the source and destination tasks of the messages. It is necessary to include the periods of the corresponding global messages π_m, i.e. the periods of their source tasks, and their destination tasks, because in the worst case the source task misses the first instance of a message transmission, and the second instance arrives immediately after the start of a destination task, thus will not be read until the next instance of this task, as depicted in Fig. 28.5.

According to the model of Zhu et al. [18] we expand the master problem (28.1) by

$$\begin{aligned}\sum_{s\in S}\ \sum_{p\,\in\, P^s\ :\ t\,\in\, p\ \cap\ c} r_{t,p}\cdot X_{s,p} + \sum_{m\,\in\, q\ :\ m_a\,\in\, c\ \wedge m_B\,\in\, c} \pi_m\cdot Y_m \\ + \sum_{q\in Q}\ \sum_{m\,\in\, q\ :\ m_a\,\in\, c\ \wedge m_B\,\in\, c} (r_{m,q}+\pi_{m_B})\cdot Z_q \le d_c\end{aligned} \tag{28.23}$$

for all task chains $c \in C$, where C denotes the set of all task chains, and their corresponding end-to-end deadlines d_c. Another artificial variable is needed to keep the master problem feasible until the WCRTs of the tasks and messages in the chain can satisfy (28.23).

In order to manage task chains the presented way we need to compute the exact WCRTs of the tasks and messages, and therefore we have to use the scheduling approach of Eisenbrand et al. and cannot use the DMS scheduling approach of Lehoczky et al. presented in Sect. 28.3.3.

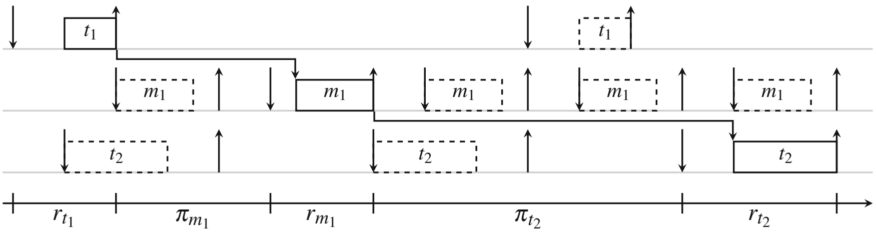


Fig. 28.5 An example of the worst-case end-to-end latency occurring in the transmission of a message $m_1 = (t_1, t_2)$ over the global bus. Task t_1 misses the first instance of message m_1, and the second instance of m_1 finishes shortly after the start of task t_2. The down arrows indicate the activation times, the up arrows the WCRTs of tasks and messages

28.3.5 The Global FlexRay Bus

There are three common bus systems that can be considered for the global communication: first, a token area network (TAN) bus, in which a token is given to the subsystems in a round-robin fashion and the gateway ECU holding the token can send over the bus as discussed by Althaus et al. [2]. Second, a controller area network (CAN) bus, in which the messages gain a priority for a non-preemptive transmission by Davis et al. [8]. Third, a FlexRay bus consisting of a deterministic static segment resembling a time-division multiplexing access (TDMA) fashion, which is analyzed by Lukasiewycz [14], and a CAN bus-like dynamic segment, which we neglect as it is non-deterministic.

In a first underapproximation model we only ensure that the number of static slots is sufficient for every message transmitted unaccompanied by extension of the message pricing problem (28.15). For the FlexRay bus we highly abstract and assume that every slot is sufficiently dimensioned for each message, and that multiplexing is not used. Then we compute a lower bound on the signal's response time r_m in a straightforward manner by bounding it with the maximum time distance between allocated slots for this message.

The local transmission within a subsystem and within the ECUs is neglected in both implemented approaches.

28.4 Experiments

In [1] we presented the proof-of-concept experiment as described in Sect. 28.4.1. In this work we additionally present a benchmark arises in the context of a Virtual Driver Assistance (ViDAs) system, see Sect. 28.4.2.

28.4.1 Proof-of-Concept Experiment

As an example, we use a synthetic architecture of two subsystems with at most three ECUs of two different ECU types of cost 10 and 40 in each subsystem, and two copies of the task network of Fig. 28.3 with the given WCETs on the ECU type of cost 10 and the halved WCETs on the ECU type of cost 40. Additionally, we created three signals and one taskchain for a global FlexRay bus of two slots, where one slot is already occupied to transmit one signal from two predeployed tasks to different subsystems.

The optimal allocation in every case uses three ECUs of the cheap ECU type with total cost of 30. More realistic constraints, e.g. that a group of tasks has to be executed on the same ECU, are not activated although they are already implemented.

The experiments were executed on one core of a server with two six-core processors (Intel$^{\circledR}$ Core$^{\text{TM}}$ i7–970 at 3,2 GHz) with 12 GB RAM. For solving the generated LPs and ILPs we use the commercial LP solver Gurobi Optimizer 5.0.2 [10].

In Table 28.1 we show the values and running times of the first and best lower bounds obtained by different approaches. For the example without a task chain we used our LP formulation of the approach of Lehoczky et al. [12]. If we consider task chains with end-to-end deadlines, we have to compute the exact response times, and thus use a reimplementation of the approach of Eisenbrand et al. [9]. The heuristic reference value is provided by the approach of Thaden et al. [16].

The approach is able to handle different degrees of underapproximation to fast provide lower bounds. In the memory case we only ensure the schedulability of the periodic tasks and their memory consumption. For the global message transmission we first give the results for the described simple slot control, then our abstraction of the FlexRay bus.

28.4.1.1 Evaluation

As expected the quality of the lower bounds increases with the degree of the model and the runtime. The fact that the weaker model is solved slower will be

Table 28.1 Results of the different approaches

Scheduling model	Without task chains		With task chains	
	First LB	Best LB	First LB	Best LB
Memory	24 1.0s	30 3.4s	28 5.3s	30 29.6s
+ Slot control	24 0.6s	30 2.6s	28 4.8s	30 24.1s
+ FlexRay bus	24 0.9s	30 3.1s	28 4.9s	30 24.2s
Heuristic	30 0.2s		30 0.2s	
Optimal value	30		30	

Table 28.2 Hardware architecture of one of 5 identical subsystems in the ViDAs benchmark

ECU	leon3_cache cost = 32	leon3_no_cache cost = 28	arm7_std cost = 22	arm7_fast cost = 25
ECU0☆	avail	avail	avail	preset
ECU1	avail	avail	avail	avail
ECU2	avail	avail	avail	avail
ECU3	avail	avail	avail	avail

☆ Gateway ECU, i.e. connection to the global bus.

compensable for larger instances as the ILPs will get larger. In this case a good lower bound is already expectable with a weaker model in less time.

The exact computation of the response times is very time consuming considering task chains, because of the more complex ILP formulation in the pricing problems.

28.4.2 ViDAs Benchmark

In this work we present a benchmark arising from a student project of the University of Oldenburg. The task was to develop a Virtual Driver Assistance (ViDAs) system capable of automatically joining a two-lane motorway by using sensors to identify a sufficiently large gap between cars driving in the right lane.

To obtain a realistic hardware architecture, we refer to the academic benchmark from Tindell et al. [17] which we extend to 5 identical subsystems and 4 ECU types, see Table 28.2, and we generate 5 aggregated tasks that work as a base load, see Table 28.3. For that reason the mapping of the base load tasks is preset to a specific ECU type which leads to initial costs of 125.

The additional tasks are obtained by a task creation process as described in Büker [4] where we ignore task chains, so that we can use the approach described in Sect. 28.3.3. This process partitions the ViDAs software into multiple software tasks and signals that we copied to obtain larger networks, see Table 28.3.

The experiments were executed on 12 cores, of a server with four 16-core processors (AMD Opteron$^{\text{TM}}$ 6282 SE at 2,6 GHz) with a timeout of 2 hours. For solving the generated LPs and ILPs we use the commercial LP solver Gurobi Optimizer 5.0.2 [10], resp. IBM ILOG CPLEX 12.4.

In Fig. 28.6 we show the determined cost and runtime of different approaches with a simple slot control for the global bus. The heuristic approaches KL, LP_GRB and LP_CPX are described in Thaden [15], CGS illustrates different kinds of our column generation approach of Sect. 28.3.3. The first lower bound (CGS_1LB) is already obtained after solving the master LP for the first time, i.e. it must not relate to an integral solution, instead of the first integral solution (CGS_INT) that is generated by a special branching routine and therefore can be seen as a heuristic solution after checking the feasibility of the local bus, which is

Table 28.3 Properties of the ViDAs benchmark

Task /	Mapping		leon3_cache	leon3_no_cache	arm7_std	arm7_fast	deadline	period
Signal	SUB	ECU	WCET	WCET	WCET	WCET		
TASK0	SUB0	ECU0	480	640	1600	1530	4000	4000
TASK1	SUB1	ECU0	720	960	3300	2940	4000	4000
TASK2	SUB2	ECU0	1200	1600	–	6000	6000	6000
TASK3	SUB3	ECU0	1200	1600	–	6000	6000	6000
TASK4	SUB4	ECU0	1200	1600	–	6000	6000	6000
TASK10			2505	2706	2751	2813	5000	5000
TASK11			2007	2007	2004	2005	5000	5000
TASK12			2029	2030	2018	2022	5000	5000
TASK13			2611	2844	2925	3003	5000	5000
TASK14			2098	2123	2112	2124	5000	5000
TASK15			2403	2455	2175	2205	5000	5000
TASK16			2261	2284	2481	2510	5000	5000
SIGNAL0	(TASK16 → TASK10)							
SIGNAL1	(TASK13 → TASK12)							
SIGNAL2	(TASK15 → TASK14)							
SIGNAL3	(TASK13 → TASK15)							

5 base load tasks (first block), 7 additional tasks (second block), and 4 additional signals (third block) that will always fit into one of 26 bus slots.

neglected in this approach. The best lower bound (CGS_bLB) is determined after solving the problem instance to its optimum.

Instance INST1 consists of the 5 base load tasks and one copy of 7 additional tasks with 4 additional signals. Instance INST2 consists of the 5 base load tasks and two copies of this network, whereas instance INST3 consists of 3 copies.

28.4.2.1 Evaluation

We provide safe lower bounds to all instances of the ViDAs benchmark in comparable time. As expected the runtime of our algorithm excessively increases with the size of the task network. The fact that the computation of the first integral solution needs more time than the best lower bound is due to a different branching strategy in this case. If we neglect the local transmission, it is also observable that every integral solution of our approach can be seen as a heuristic solution that is computed more stable than in the case of the ILP-based global analyses that run out of time.

In more realistic scenarios it can be expected, that a solution has to fulfill more constraints, e.g. several tasks have to run or are forbidden to run on the same ECU or in the same subsystem, that is already applicable to our approach and will reduce the search space of the pricing problems tremendously.

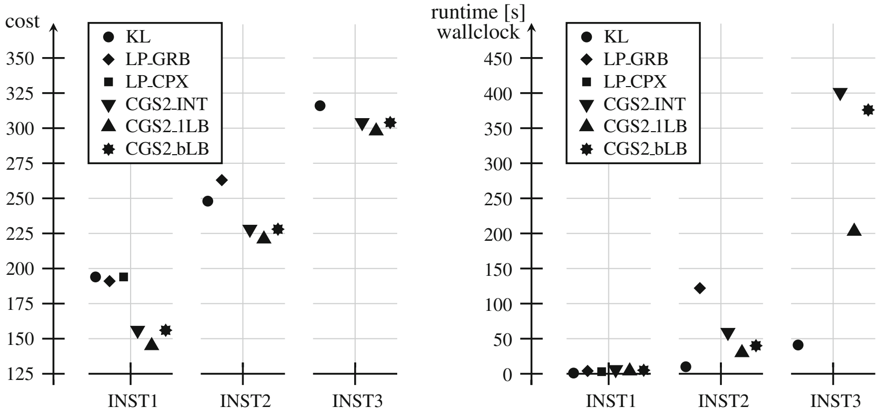


Fig. 28.6 Determined cost (*left*) and runtime (*right*) of the ViDAs benchmark for different approaches. KL: Global analysis based on a Kernighan-Lin partitioning algorithm; LP_GRB: Global analysis with ECU type bins based on an ILP with solver Gurobi, LP_CPX: Same analysis with solver CPLEX; CGS_INT: First integral solution, CGS_1LB: First lower bound, CGS_bLB: Best lower bound by our column generation scheduling approach with solver Gurobi

28.5 Conclusion

We presented a column generation approach for scheduling of real-time networks which satisfies constraints concerning the subsystems, tasks, ECU types, ECUs and messages as well as task chains. This approach extends the approach of Althaus et al. [2] in the sense of a wider perspective of the real-time network.

Our approach can be integrated in a hybrid algorithm that smartly finds heuristic solutions while proving their quality, and in the best case their optimality. A speed-up of the solving process is expected if we use heuristic solutions as start columns for our approach.

For solving larger task networks more efficiently the number of variables and constraints in the pricing problems has to be reduced. We have in mind solving easier, non-exact relaxations of the pricing problems in a combinatorial way, and verify the results appropriately to improve the search for new variables in the column generation approach, provided that we pinpoint which constraints are responsible for the hardness of this problem.

Acknowledgments This work was partly supported by the German Research Council (DFG) as part of the Transregional Collaborative Research Center *Automatic Verification and Analysis of Complex Systems* (SFB/TR 14 AVACS, www.avacs.org).

References

1. Althaus E, Hoffmann S, Kupilas J, Thaden E (2012) A column generation approach to scheduling of real-time networks. In: Proceedings of the world congress on engineering and computer science (WCECS), vol 1, IAENG, San Francisco, USA, pp 224–229
2. Althaus E, Naujoks R, Thaden E (2011) A column generation approach to scheduling of periodic tasks. In: Experimental slgorithms—10th international symposium, SEA 2011, Proceedings, LNCS 6630, vol 1, Springer, Berlin, pp 340–351
3. Audsley NC (1990) Deadline monotonic scheduling
4. Büker M (2012) An automated semantic-based approach for creating task structures. Ph.D. thesis
5. Büker M, Damm W, Ehmen G, Metzner A, Stierand I, Thaden E (2011) Automating the design flow for distributed embedded automotive applications: keeping your time promises, and optimizing costs, too. In: Proceedings international symposium on industrial embedded systems (SIES'11)
6. Burns A, Wellings A (2001) Real-time systems and programming languages: Ada 95, real-time Java, and real-time POSIX. Addison-Wesley, International computer science series, Reading
7. Clark B, Stierand I, Thaden E (2011) Cost-minimal pre-allocation of software tasks under real-time constraints. In: Proceedings of the 2011 ACM symposium on research in applied computation (RACS 2011), Miami, Florida, pp 77–83
8. Davis RI, Burns A, Bril RJ, Lukkien JJ (2007) Controller area network (CAN) schedulability analysis: Refuted, revisited and revised. Real-Time Syst 35(3):239–272
9. Eisenbrand F, Damm W, Metzner A, Shmonin G, Wilhelm R, Winkel S (2006) Mapping task-graphs on distributed ecu networks: efficient algorithms for feasibility and optimality. In: Proceedings of the 12th IEEE conference on embedded and real-time computing systems and applications. IEEE Computer Society
10. Gurobi Optimization, Inc. (2012) Gurobi optimizer reference manual (2012).http://www.gurobi.com
11. Joseph M, Pandya PK (1986) Finding response times in a real-time system. Comput J 29:390–395
12. Lehoczky JP, Sha L, Ding Y (1989) The rate monotonic scheduling algorithm: exact characterization and average case behavior. In: IEEE Real-time systems, symposium, pp 166–171
13. Leung JYT, Whitehead J (1982) On the complexity of fixed-priority scheduling of periodic, real-time tasks. Perform Eval 2(4):237–250
14. Lukasiewycz M, Glaß M, Teich J, Milbredt P (2009) FlexRay schedule optimization of the static segment. In: CODES+ISSS, ACM, pp 363–372
15. Thaden E (2013) Semi-automatic optimization of hardware architectures in embedded systems. Ph.D. thesis
16. Thaden E, Lipskoch H, Metzner A, Stierand I (2010) Exploiting gaps in fixed-priority preemptive schedules for task insertion. In: Proceedings of the 16th international conference on embedded and real-time computing systems and applications (RTCSA), (IEEE) Computer Society, pp 212–217
17. Tindell K, Burns A, Wellings A (1992) Allocating hard real time tasks (an NPhard problem made easy). J Real-Time Syst 4:145–165
18. Zhu Q, Yang Y, Natale MD, Scholte E, Sangiovanni-Vincentelli AL (2010) Optimizing the software architecture for extensibility in hard real-time distributed systems. IEEE Trans Industr Inf 6(4):621–636

Chapter 29
A Unifying Framework for Parallel Computing

Victor Eijkhout

Abstract We propose a new theoretical model for parallelism. The model is explictly based on data and work distributions, a feature missing from other theoretical models. The major theoretic result is that data movement can then be derived by formal reasoning. While the model has an immediate interpretation in distributed memory parallelism, we show that it can also accomodate shared memory and hybrid architectures such as clusters with accelerators.The model gives rise in a natural way to objects appearing in widely different parallel programming systems such as the PETSc library or the Quark task scheduler. Thus we argue that the model offers the prospect of a high productivity programming system that can be compiled down to proven high-performance environments.

Keywords Dataflow · Data distribution · High performance computing · Hybrid architecture · Message passing · Parallel programming

29.1 Introduction

As computer architectures become larger in scale and more sophisticated in their hybrid nature (cluster, shared memory, accelerators), the problem of high productivity high performance programming is becoming acute. The problem is only to a limited extent one of the low level programming models: the major part of the problem is the parallel coordination of cores, devices, cluster nodes, co-processors, et cetera.

Solutions such as CUDA or MPI have any number of limitations, foremost among which that they are all special purpose, so it is not possible to write a code that is portable between systems. Also, such programming systems are often of a

V. Eijkhout (✉)
Texas Advanced Computing Center (TACC), The University of Texas at Austin,
Austin, TX, USA
e-mail: eijkhout@tacc.utexas.edu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_29,

low level, asking the programmer to be concerned with details that are not essential to the application.

In this paper we give the design of a system that takes an abstract approach to implementing parallel algorithms where specific architectural details can be explicitly modeled. Our Integrative Model for Parallelism (IMP) (earlier presented in [6]) allows for an abstract description of an algorithm, that can be made explicit through successive transformations, one of which being its convolution with an abstract description of hardware.

Specifically, the IMP model is based on kernels, which correspond to parallel tasks without data dependencies. From the distribution of work and data, data movement is theoretically derived as a first-class object. A single kernel typically corresponds to a collective operation, or a bulk data transfer such as exists in the PETSc [10] and Trilinos [11] libraries.

By composing IMP kernels we arrive at an ‘abstract algorithm’, which is takes the form of a Directed Acyclic Graph (DAG) or dataflow diagram. Actual data motion results from an assignment of tasks to ‘computing locales’: threads, cores, nodes, et cetera. The strength of the IMP model is that this assignment is again an explicit feature of the model, so data motion becomes derivable from the abstract algorithm, rather than being explicitly coded as MPI messages, or implicitly resulting as side-effect of the execution. We will explain this in detail, and give motivating examples.

Having an explicit data motion object has several advantages: for one, it means that a programmer does not have to code in terms of send/receive. For another, the same communication pattern is often used several times in a row. Thus, having a data motion object allows for any preprocessing for optimizing the communication schedule to be amortized. This is known as the ‘inspector-executor’ model [22].

A few things our model is not. It is a programming model, so we offer no transformations of existing codes. It is not a cost model, though cost can be included in our formal derivations. We do not propose a new programming language: we feel that high performance can be reached by compiling down to already existing tools. We do not claim to be able to derive optimal algorithms, routing, or scheduling: the programmer still has the responsibility for the algorithm design; we offer a high level, high productivity way of expressing the design.

29.2 Survey of Earlier Models

We briefly survey a number of directions in parallel programming research.

29.2.1 Distributed Memory Systems

Distributed memory programming is mostly characterized by being the *de facto* standard on large systems, and having been mostly ignored by theory. We survey the dominant ideas.

29.2.1.1 Message Passing

Starting around 1990 a great many software packages were developed that, with high performance in mind, systematized communication (especially for distributed-memory systems), but put a considerable burden on the programmer. Foremost among these is the Message Passing Interface (MPI) [9]. While MPI solely used a two-sided communication model (until the MPI-2 standard was formalized), around the same time several one-sided models were developed, such as *shmem* [5] and Charm++ [12]. The latter package offers "active messages", which are an important generalization of plain data transfer: in addition to being one-sided, they associate operations with the transferred data.

One problem with MPI is that it is too explicit: there is no level on which a software layer 'understands' the communication, for instance in order to transform and optimize it.

29.2.1.2 Inspector Corrector

Another HPC system, which in fact predates MPI, is the "Parti primitives" [22]. This package originated the inspector-executor model, where the user would first declare a communication pattern to an inspector routine, which would then yield an object whose instantiation was the actual communication. This expresses the fact that, in High Performance Computing (HPC), communications are often repetitions of the same irregular pattern. The inspector-executor model is currently available in the `VecScatter` object of the PETSc library [10], and the `map` objects of Trilinos [11].

29.2.1.3 Bulk Synchronous Parallelism

BSP is another conceptual model (the author talks about a 'bridging model') for describing parallelism [23]. This model uses asynchronous one-sided messages and 'supersteps' that use barrier synchronization to resolve the asynchronicity. As we will see, a single application of a kernel in our model is comparable to a superstep. However, IMP kernels do not have an explicit barrier synchronization point, and in fact a sequence of kernels can execute completely asynchronously.

29.2.2 Dataflow

Dataflow is the notion that an execution unit can start ('fire') when all of its input are available. Attempts to use this on the level of instruction level parallelism have largely failed because of overhead. However, a number of programming systems have applied this to task-sized execution units [1, 13]. In the context of parallel processing, some authors have extended dataflow to the case where a task is spawned in parallel on multiple processors [17]. Such systems often implicitly assume a shared memory runtime layer, where arbitrary processes can be created.

29.2.3 Global Address Space

There are a number of systems that implicitly or explicitly make the assumption that any processing elements can access any data element.

29.2.3.1 PGAS Languages

Under the DARPA HPCS program several groups have tried to devise programming languages that alleviate the programmer's burden, while not entirely abstracting the architecture into nothingness. Probably most interesting among these is Chapel [4], which has elaborate mechanisms for indicating groups of processors and the distribution of data and work among them. Other PGAS languages in use are UPC and X10.

29.2.3.2 Programming Systems

In recent years, a number of systems have been developed that address less regular parallel applications than are traditionally handled by MPI. We mention ParalleX [7], Galois [15], and Chorus [16]. Such systems offer elegant descriptions of irregular applications such as graph algorithms; however, they do so at the price of introducing a middle software layer that makes distributed memory act like shared. It is unclear whether this approach will scale up to the current generation of 100k core machines and beyond.

One successful package, the "Global Arrays" package [18], implements arrays that act as if they are in shared memory; it uses BSP type synchronization.

29.2.3.3 DAG Schedulers

One particular type of dataflow is DAG based, encountered in some contemporary packages that were primarily developed for linear algebra. Packages such as Quark [24] and SuperMatrix [3, 20] handle this by declaring tasks that have multiple areas of memory as input and output, essentially realizing a dataflow dependency implementation. Other packages in this vein are Intel CnC http://software.intel.com/en-us/articles/intel-concurrent-collections-for-cc, SMPSs, StarPU, Uppsala.

29.3 The Basic I/MP Model

In this section we develop the basic theoretical framework of the IMP model. We introduce the concept of a distributed kernel, and show how it allows for formal derivation of data movement.

29.3.1 Basics

We define a *kernel* in the IMP model as a directed bipartite graph, that is, a tuple comprising an input data set, an ouput data set, and a set of edges denoting elementary computations that take input items and map them to output items:

$$K = \langle \mathrm{In}, \mathrm{Out}, E \rangle$$

where

In, Out are data structures, and E is a set of (α, β) elementary computations, where $\alpha \in \mathrm{In}$, $\beta \in \mathrm{Out}$,

$$\mathrm{In}, \mathrm{Out} \quad \text{are} \quad \text{data} \quad \text{structures,} \quad \text{and}$$

$$E \text{ is a set of } (\alpha, \beta) \text{ elementary computations, where } \alpha \in \mathrm{In}, \beta \in \mathrm{Out},$$

and ‘elementary computations’ are simple computations between a single input and output. Multiple edges reaching a single output element is interpreted as a reduction operation. For now we can assume that a rule for resolving such multiple accesses is in place.

To parallelize a kernel over P processors, we define

$$K = \langle K_1, \ldots, K_P \rangle, \qquad K_p = \left\langle \mathrm{In}_p, \mathrm{Out}_p, E_p \right\rangle,$$

describing the parts of the input and output data set and (crucially!) the work that are assigned to processor p. The only restrictions on these distributions are

$$\mathrm{In} = \bigcup_p \mathrm{In}_p, \mathrm{Out} = \bigcup_p \mathrm{Out}_p, E = \bigcup_p E_p;$$

none of these distributions are required to be disjoint. To foreshadow the rest of the discussion in this section, we remark that elementary computations in E_p (meaning that they are executed on processor p) need not have their input data in In_p, nor their output in Out_p. The communication in a parallel algorithm will be seen to follow directly from the relations in processor locality between input/output data sets and elementary computations.

29.3.2 Derivation of Data Movement

Based on the fact that the computations in E_p are executed on processor p we can now define the input and output data for these computations:

$$\begin{array}{rl} \mathrm{In}(E_p) & \{\alpha : (\alpha, \beta) \in E_p\}, \\ \mathrm{Out}(E_p) & \{\beta : (\alpha, \beta) \in E_p\}. \end{array}$$

These correspond to the input elements that are needed for the computations on processor p, and the output elements that are produced by those computations. These sets are related to $\mathrm{In}_p, \mathrm{Out}_p$ but are not identical: in fact we can now characterize the communication involved in an algorithm as

$$\begin{cases} \mathrm{In}(E_p) - \mathrm{In}_p & \text{data to be communicated to p before computation on p} \\ \mathrm{Out}(E_p) - \mathrm{Out}_p & \text{data computed on p to be communicated out afterwards.} \end{cases}$$

We see that some simple cases are covered by our model: the common ‘owner computes’ case corresponds to

$$\mathrm{Out}(E_p) = \mathrm{Out}_p,$$

that is, each processor computes the elements of its part of the output data structure and no data is communicated after being computed. If additionally $\mathrm{In}(E_p) = \mathrm{In}_p$, we have an embarrassingly parallel computation because no communication occurs before or after computation.

29.3.3 From Kernels to Algorithms

Next we consider constructing a full algorithm by composing kernels. This gives us an ‘abstract algorithm’: a description of data dependencies without regard for architectural details.

This composition process can be interpreted as graph construction, or as function composition if we interpret kernels in a functional manner; $\mathrm{Out} \leftarrow K(\mathrm{In})$.

29.4 Distributions

So far we have explained the IMP model in the abstract. We will now present distributions: a construct that can be basis for an IMP-based programming system.

Informally, a distribution is a mapping from processors to data. In this sense, each kernel features on input the distributions $p \mapsto \mathrm{In}_p$ (which is determined by the context, and typically disjoint), and $p \mapsto \mathrm{In}(E_p)$ (which is determined by the kernel and typically not disjoint). Similarly, $p \mapsto \mathrm{Out}_p$ and $p \mapsto \mathrm{Out}(E_p)$ are the distributions at the output of a kernel. Thus, we see that the communication prior to and after a kernel can be formulated as the transformation from the context distribution to the kernel distribution and vice versa.

29.4.1 Basic Definition

Let us consider a vector[1] of size N and P processors. A distribution is a function that maps each processor to a subset of N[2]:

$$v : P \to 2^N.$$

Thus, each processor stores elements of the vector; the partitioning does not need to be disjoint.

We point out as a special case the redundant replication

$$* \equiv p \mapsto N.$$

Let x be a vector and v a distribution, then we can introduce an elegant, though perhaps initially confusing, notation for distributed vectors[3]:

$$x(v) \equiv p \mapsto x[v(p)] = \{x_i : i \in v(p)\}.$$

That is, $x(v)$ is a function that gives for each processor p the elements of x that are stored on p according to the distribution v. As an important special case, $x(*)$ describes the case where each processor stores the whole vector.

We can simply extend the concept of vector distributions to matrix distributions. First we consider matrices that are distributed by block rows or columns. If u is a vector distribution, we can define

[1] We can argue that this is no limitation, as any object will have a linearization of some sort.

[2] We make the common dentification of $N = \{0, \ldots, N-1\}$ and $P = \{0, \ldots, P-1\}$; likewise N^M is the set of mappings from M to N, and thereby 2^N is the set of mappings from N to $\{0, 1\}$; in effect the set of all subsets of $\{0, \ldots, N-1\}$.

[3] We use parentheses for indicating distributions; actual vector subsections are denoted with square brackets.

$$A(u,*) \equiv p \mapsto A[u(p),*]$$
$$A(*,u) \equiv p \mapsto A[*,u(p)]$$

As observed above, the most important application of distributions is converting a vector between one distribution and another. We use the notation $T(u,v)$ for this conversion, so

$$x(v) = T(u,v)x(u). \tag{29.1}$$

This transformation corresponds to the `VecScatter` object in PETSc or the `map` object in Trilinos. Thus, we have shown how our model derives a construct underlying code in two highly successful ($> 100,000$ cores) scientific libraries.

29.4.2 Example: Matrix-Vector Product

The dense matrix-vector product provides a simple illustration of the principle that a parallel algorithm can be written in terms of distributions, and that data movement is formally derived. We make our point by giving two versions of the algorithm, and showing how the resulting very different data traffic patterns follow automatically.

We write the basic operation $\forall_i : y_i = \sum_j a_{ij}x_j$ by splitting the computation into temporaries and reduction:

$$\begin{aligned}\forall_i : y_i &= \sum\nolimits_j a_{ij}x_j\\ &= s\sum\nolimits_j t_{ij}, \quad t_{ij} = a_{ij}x_j.\end{aligned}$$

For simplicity we assume a square matrix, and the same distribution u is used for both input and output, that is, we assume that x is distributed as $x(u)$ on input, and y as $y(u)$ on output. The work distribution is determined by the decision to let t_{ij} be computed where a_{ij} is stored. We will now derive communication for the variants of the product operation induced by the matrix distribution by rows and columns.

Product by rows

We write the product by rows as

$$\begin{cases} t(u,*) \leftarrow A(u,*) \cdot_\times x(*)\\ \quad y(u) \leftarrow \sum_j t(u,*). \end{cases}$$

where dot-times indicates a column scaling by right multiplication, and we reason as follows:

- $A(u,*)$ describes the distribution of A by block rows.

- In order to perform the product, x needs to be distributed as $x(*)$. Since x is distributed as $x(u)$, the transformation to $x(*)$ is an allgather.
- The temporaries are distributed upon construction as $t(u,*)$, which is the correct distribution for the reduction, so no communication is needed there.
- On output, y is distributed as $y(u)$, which is the desired distribution, so no further communication is needed.

Product by columns
Similarly, we write the matrix-vector product by columns in distribution terms as

$$\begin{cases} t(*,u) \leftarrow A(*,u) \cdot_{\times} x(u) \\ y(u) \leftarrow \sum_j t(u,*) \end{cases}$$

The reasoning is now:

- $A(*,u)$ describes the distribution of A by columns.
- x is distributed as $x(u)$, which is the required distribution for the t_{ij} calculation.
- $t(*,u)$ is not correctly distributed for the reduction, so a data transposition is needed here.
- $y(u)$ is the resulting reduction as before.

We note that the transpose and subsequent reduction can be merged into reduce-scatter operation. Our framework does not derive such a merge, rather we assume that software tools can effect this [19].

In these examples we see the essence of the IMP model: the algorithm is expressed in global terms, and the actual data movement follows from the data and work distributions. In particular, it is not explicitly coded. Deriving the communication pattern can be done in a library or by a relatively simple code transformation tool. By noting the resemblance of our derived data movement patterns to existing high performance library constructs, we hope to have argued that our model can be a high performance programming environment.

29.4.3 Kernels and Distributions

We associate with every kernel $K = \langle \mathrm{In}, \mathrm{Out}, E \rangle$ four distributions.

- The α-distribution of a kernel is $p \mapsto \mathrm{In}_p$;
- The β-distribution is $p \mapsto \mathrm{In}(E_p)$;
- The δ-distribution is $p \mapsto \mathrm{Out}(E_p)$; and
- The ϵ-distribution is $p \mapsto \mathrm{Out}_p$.

Thus, every kernel consists of the $T(\alpha, \beta)$ traffic, followed by local computation, followed by the $T(\delta, \epsilon)$ traffic.

29.5 Example: N-body Problems

We will show a nontrivial examples of how algorithms are expressed and analyzed in the IMP model, using both kernels and distributions.

Algorithms for the N-body problem need to compute in each time step the mutual interaction of each pair out of N particles, giving an $O(N^2)$ method. However, by suitable approximation of the 'far field' it becomes possible to have an $O(N \log N)$ or even an $O(N)$ algorithm, see the Barnes-Hut octree method [2] and the Greengard-Rokhlin fast multipole method [8].

The naive way of coding these algorithms uses a form where each particle needs to be able to read values of in principle every cell. This is easily implemented with shared memory or an emulation of it.

However, this algorithm can be implemented just as easily in distributed memory, using message passing [21]. To show that an implementation can be formally derived we consider the following form of the N-body algorithms (see [14]):

- The field due to cell i on level ℓ is given by

  $$g(\ell, i) = \oplus_{j \in C(i)} g(\ell + 1, j)$$

  where $C(i)$ denotes the set of children of cell i and $\oplus$ stands for a general combining operator, for instance computing a joint mass and center of mass;
- The field felt by cell i on level ℓ is given by

  $$f(\ell, i) = f(\ell - 1, p(i)) + \sum_{j \in I_\ell(i)} g(\ell, j)$$

where $p(i)$ is the parent cell of i, and $I_\ell(i)$ is the interaction region of i: those cells on the same level ('cousins') for which we sum the field.

29.5.1 *Kernel Implementation*

We can model the above formulation straightforwardly in terms of IMP kernels: the g computation has $E^{(g)} = E^\tau \cup E^\gamma$, where

$$\begin{cases} E^{\tau} = \{\tau_{ij}^{\ell}\} \\ E^{\gamma} = \{\gamma_i^{\ell}\} \end{cases}, \begin{cases} \forall_i \forall_{j \in C(i)} : & \tau_{ij}^{\ell} = \text{` } t_{ij}^{\ell} = g_j^{\ell+1}\text{'} \\ \forall_i : & \gamma_i^{\ell} = \text{` } g_i^{\ell} = \oplus_{j(i)} t_{ij}^{\ell}\text{'} \end{cases} \tag{29.2}$$

The t_{ij}^{ℓ} quantities are introduced so that their assignment can model data communication: as in the matrix-vector example above, the g_i^{ℓ} reduction computation is then fully local.

Similarly, the f computation is $E^{(f)} = E^{\rho} \cup E^{\sigma} \cup E^{\phi} \cup E^{\eta}$, where

$$\begin{cases} E^{\rho} = \{\rho_i^{\ell}\} \\ E^{\sigma} = \{\sigma_{ij}^{\ell}\} \\ E^{\eta} = \{\eta_i^{\ell}\} \\ E^{\phi} = \{\phi_i^{\ell}\} \end{cases}, \begin{cases} \forall_i : & \rho_i^{\ell} = \text{` } r_i^{\ell} = f_{p(i)}^{\ell-1}\text{'} \\ \forall_i \forall_{j \in I_{\ell}(i)} : & \sigma_{ij}^{\ell} = \text{` } s_{ij}^{\ell} = g_j^{\ell}\text{'} \\ \forall_i : & \eta_i^{\ell} = \text{` } h_i^{\ell} = \sum_{j \in I_{\ell}(i)} s_{ij}^{\ell}\text{'} \\ \forall_i : & \phi_i^{\ell} = \text{` } f_i^{\ell} = r_i^{\ell} + h_i^{\ell}\text{'} \end{cases} \tag{29.3}$$

These formulations can immediately be translated to a message passing implementation.

Transformations of the algorithm are possible. For instance, in the statement

$$\forall_i \forall_{j \in I_{\ell}(i)} : s_{ij}^{\ell} = g_j^{\ell}$$

we recognize a broadcast of g_j^{ℓ} to all nodes i such that $j \in I_{\ell}(i)$. We can reformulate it as such by exchanging the quantifiers:

$$\forall_j \forall_{i \in J_{\ell}(j)} : s_{ij}^{\ell} = g_j^{\ell}$$

where

$$J_{\ell}(i) = \{j : j \in I_{\ell}(i)\}.$$

29.5.2 Practical Aspects

We have here given an implementation of tree algorithms for N-body problems. This implementation can cope with the difficulties that distributed memory imposes; as indicated above, shared memory implementations are considerably easier to describe. However, our implementation essentially gives a dependency graph of tasks, hence can also serve as shared memory implementations.

In the distributed memory case we invoke the inspector-executor paradigm (see the introduction): we determine which elements need to communicate, in particular the $I_{\ell}(i)$ and $C_{\ell}(i)$ sets, and use this information to evaluate irregular gathers repeatedly.

29.6 Derivation of Physical Data Movement

The model of sect. 29.3 was built around the concepts of 'kernel' and 'abstract algorithm'. Algorithms were purely formulated in terms of dataflow and data dependencies; in particular no architectural considerations were taken into account. This model implicitly uses a flat processor structure: all processes can send to and receive from all others and all communications are treated equally. Thus, the model has direct applicability to distributed memory parallel computing with an interconnect that is essentially all-to-all, such as a fat-tree.

However, in other circumstances it fails to account for several aspects:

- A distributed memory architecture can have a mesh interconnect, or other scheme where certain processor pairs do not have a direct connection.
- As a special case of a not-fully-connected network we may consider a cluster with accelerators: the accelerators do not connect to the network but only to a host processor.
- The model does not suggest any scheduling for operations, and the available parallelism can greatly exceed the number of physically available processors.
- In shared memory, using threading, several processes live in the same physical address space. In this case, certain data dependencies correspond to a data movement no-op; also, the precise timing of tasks then becomes an issue.

In order to cover these aspects we need to include the mapping of tasks to computing loci in the IMP model.

Clusters with shared memory nodes

We consider a common example of distributed memory clusters where each node supports multiple processes that live on shared memory. This is illustrated in Fig. 29.1, where we have two cluster nodes with two shared memory processes each. In this case, algorithm edges $(1,2)$ or $(3,4)$ do not need an MPI message, while edges such as $(1,4)$ do.

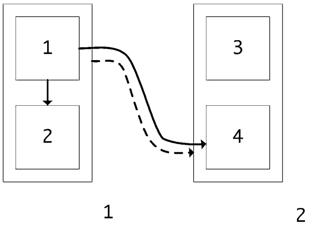


Fig. 29.1 Communication between processes (*solid*) and nodes (*dotted*) in a cluster/shared memory architecture

Thus, we have two graphs: the process graph from the IMP kernel, connecting four processes, and the resulting *processor* graph, connecting two processors. We derive the latter from the former by a linear transformation. Let us consider specifically the IMP dependencies in Fig. 29.1.

First we form the embedding operator from the processes to the cluster nodes

$$I_2^4 = \begin{pmatrix} \star & \star & \cdot & \cdot \\ \cdot & \cdot & \star & \star \end{pmatrix}$$

that reflects that the first two processes live on node 1, while the third and fourth live on node 2. We denote its transpose by I_4^2.

The process edge $(1,4)$ in the figure can now be rendered with an adjacency matrix

$$G_4 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \star & \cdot & \cdot & \cdot \end{pmatrix}.$$

First we form

$$I_2^4 \cdot G_4 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \star & \cdot & \cdot & \cdot \end{pmatrix}$$

(using the regular matrix-matrix product) to describe process behaviour in terms of nodes. If we now form the product of this with I_4^2

$$G_2 = I_2^4 \cdot G_4 \cdot I_4^2 = \begin{pmatrix} \cdot & \cdot \\ \star & \cdot \end{pmatrix}$$

we get the correct description of communications in terms of cluster nodes.

However, there is a conceptual problem with this. If we consider the intra-node edge $(1,2)$, its transform becomes

$$G_4 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \star & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix} \Rightarrow G_2 = I_2^4 \cdot G_4 \cdot I_4^2 = \begin{pmatrix} \star & \cdot \\ \cdot & \cdot \end{pmatrix}$$

stating that a message from node 1 to node 1 is required. Therefore, we introduce an extra term and form

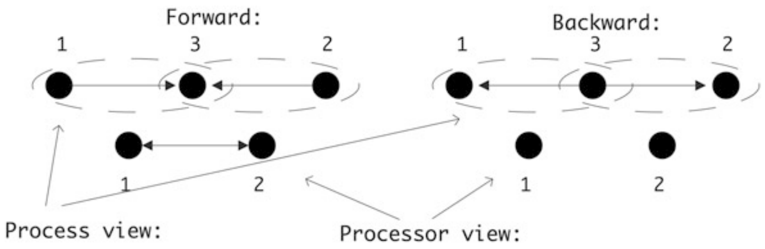


Fig. 29.2 Communication between logical and physical processors in the forward and backward sweep

$$(I_2^4 \cdot G_4) \circledast I_2^4 = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \star & \cdot & \cdot & \cdot \end{pmatrix}$$

where $A \circledast B$ is the element-by-element computation of $a \circledast b \equiv a \wedge \neg b$. The $\circledast$-multiplication by I_2^4 has the effect of limiting the communication description to only processes that are on different processors. Redoing the above examples we now find the same G_2 matrix for the inter-node case, and

$$G_2 = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix}$$

for the intra-node case, indicating that no MPI communication is needed.

Redundant processes

Next we consider redundant assignment of one process to two processors. As a specific example we take a domain decomposition method with two subdomains and one separator. In the forward sweep of the system solution the separator collects data from the subdomains; in the backward sweep it distributes data to them. We model this process by three processes, mapped to two processors, with the separator process redundantly assigned to both physical processors.

This process is illustrated in Fig. 29.2. In the left column we have the structure of the forward sweep.

The algorithmic data movement (top) takes the form of data send from the subdomains 1 and 2 to the separator 3, and the corresponding physical data movement (bottom). Since process 1 sends data to process 3, which is redundantly run on processor 2, there is a message from 1 to 2, and vice versa.

In the backward sweep, process 3 sends data to both 1 and 2, but now, since process 3 is redundantly run on both processors 1 and 2, this data movement is purely local to the processor, requiring no message passing.

We model this algebraically as follows. The forward and backward sweep are IMP kernels, which we represent by their adjacency matrices. These describe data movement between processes, that is, the algorithmic data movement:

$$L : G_3 = \begin{pmatrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \star & \star & \cdot \end{pmatrix}, \qquad U : G_3 = \begin{pmatrix} \cdot & \cdot & \star \\ \cdot & \cdot & \star \\ \cdot & \cdot & \cdot \end{pmatrix}. \tag{29.4}$$

The matrix embedding the logical processes in physical processors is

$$I_2^3 = \begin{pmatrix} \star & \cdot & \star \\ \cdot & \star & \star \end{pmatrix}$$

and by I_3^2 we denote its transpose. We can now derive the matrix of physical communications as

$$G_2 = \big((I_2^3 \cdot G_3) \circledast I_2^3\big) \cdot I_3^2$$

similar to the previous example.

If we go through this calculation for the operators in Eq. (29.4), we find

$$L : G_2 = \begin{pmatrix} \cdot & \star \\ \star & \cdot \end{pmatrix}, \qquad U : G_2 = \begin{pmatrix} \cdot & \cdot \\ \cdot & \cdot \end{pmatrix} \tag{29.5}$$

reflecting the above described behaviour that in the L sweep the processors have to exchange data, but not in the U sweep.

29.7 Conclusion

In this chapter we have presented the Integrative Model for Parallelism which offers a mode of describing parallel algorithms that is high-level and expressed in global terms. Since data movement (such as message passing in a distributed memory context) is formally derived in the model, rather than explicitly coded it offers two prospects:

- Algorithm expression is expressed independent of any particular machine model, so we can achieve portability over architecture types; and
- since data movement is derived rather than coded, we may achieve higher programmer productivity.

Additionally we have shown how architectural features can explicitly be accomodated in this model.

References

1. Adiga AK, Browne JC (1986) A graph model for parallel computations expressed in the computation structures language. In: ICPP, pp 880–886
2. Barnes J, Hut P (1986) A hierarchical O(N log N) force-calculation algorithm. Nature 324:446–449. http://dx.doi.org/10.1038/324446a0
3. Chan E, Quintana-Ortí ES, Quintana-Ortí G, van de Geijn R (2007) SuperMatrix out-of-order scheduling of matrix operations for SMP and multi-core architectures. In: SPAA '07: Proceedings of the 19th ACM symposium on parallelism in algorithms and architectures, San Diego, CA, USA, pp 116–125.
4. Chapel programming language homepage. http://chapel.cray.com/
5. Cray Research: Cray T3E^{TM} Fortran optimization guide. http://docs.cray.com/books/004-2518-002/html-004-2518-002/004-2518-002-toc.html
6. Eijkhout V (2012) A unified approach to parallel programming. In: Ao S, Douglas C, Grundfest W, Burgstone J (eds) Lecture Notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 October, 2012, San Francisco, USA, pp 78–83. Newswood Limited, International Association of Engineers. ISBN (vol I): 978-988-19251-6-9, ISBN (vol II): 978-988-19252-4-4, ISSN: 2078-0958 (Print), ISSN: 2078-0966 (Online)
7. Gao G, Sterling T, Stevens R, Hereld M, Zhu W (2007) ParalleX: a study of a new parallel computation model. In: Parallel and distributed processing symposium, 2007. IPDPS 2007. IEEE International, pp 1–6. doi:10.1109/IPDPS.2007.370484.
8. Greengard L, Rokhlin V (1987) A fast algorithm for particle simulations. J Comput Phys 73:325
9. Gropp W, Lusk E, Skjellum A (1994) Using MPI. The MIT Press, Cambridge
10. Gropp WD, Smith BF (1994) Scalable, extensible, and portable numerical libraries. In: Proceedings of the scalable parallel libraries conference, IEEE pp 87–93.
11. Heroux MA, Bartlett RA, Howle VE, Hoekstra RJ, Hu JJ, Kolda TG, Lehoucq RB, Long KR, Pawlowski RP, Phipps ET, Salinger AG, Thornquist HK, Tuminaro RS, Willenbring JM, Williams A, Stanley KS (2005) An overview of the trilinos project. ACM Trans Math Softw 31(3):397–423. http://doi.acm.org/10.1145/1089014.1089021
12. Kale LV, Krishnan S (1996) Charm++: Parallel programming with message-driven objects. In: Wilson GV, Lu P (eds) Parallel programming using C++, MIT Press, Cambridge, pp. 175–213
13. Karp RM, Miller RE (1966) Properties of a model for parallel computations: Determinacy, termination, queueing. SIAM J Appl Math 14:1390–1411
14. Katzenelson J (1989) Computational structure of the n-body problem. SIAM J Sci Stat Comput 10:787–815
15. Kulkarni M, Pingali K, Walter B, Ramanarayanan G, Bala K, Chew LP (2007) Optimistic parallelism requires abstractions. SIGPLAN Not. (Proceedings of PLDI) 42(6):211–222. http://doi.acm.org/10.1145/1273442.1250759. http://iss.ices.utexas.edu/Publications/Papers/PLDI2007.pdf
16. Lublinerman R, Chaudhuri S, Cerny P (2009) Parallel programming with object assemblies. In: International conference on object oriented programming, systems, languages and applications (OOPSLA)
17. Newton P, Browne JC (1992) The code 2.0 graphical parallel programming language. In: Proceedings of the 6th international conference on supercomputing, ICS '92, pp 167–177. ACM, New York, NY, USA. doi:10.1145/143369.143405. http://doi.acm.org/10.1145/143369.143405
18. Nieplocha J, Harrison R, Littlefield R (1996) Global arrays: A nonuniform memory access programming model for high-performance computers. J Supercomput 10:197–220
19. Poulson J, Marker B, Hammond JR, van de Geijn R Elemental: a new framework for distributed memory dense matrix computations. ACM Trans Math Softw Submitted

20. Quintana-Ortí G, Quintana-Ortí ES, van de Geijn RA, Van Zee FG, Chan E (2009) Programming matrix algorithms-by-blocks for thread-level parallelism. ACM Trans Math Softw 36(3):14:1–14:26. http://doi.acm.org/10.1145/1527286.1527288
21. Salmon JK, Warren MS, Winckelmans GS (1986) Fast parallel tree codes for gravitational and fluid dynamical n-body problems. Int J Supercomput Appl 8:129–142
22. Sussman A, Saltz J, Das R, Gupta S, Mavriplis D, Ponnusamy R (1992) Parti primitives for unstructured and block structured problems
23. Valiant LG (1990) A bridging model for parallel computation. Commun ACM 33:103–111. doi: http://doi.acm.org/10.1145/79173.79181. http://doi.acm.org/10.1145/79173.79181
24. YarKhan A, Kurzak J, Dongarra J (2011) QUARK users' guide: queueing and runtime for kernels. Technical Report ICL-UT-11-02, University of Tennessee Innovative Computing Laboratory

Chapter 30
Improving Network Intrusion Detection with Extended KDD Features

Edward Paul Guillén, Jhordany Rodríguez Parra and Rafael Vicente Paéz Mendez

Abstract In order to analyze results of anomaly detection methods for Network Intrusion Detection Systems, the DARPA KDD data set have been widely analyzed but their data are outdated for most kinds of attacks. A software called *Spleen* designed to get data from a tested network with the same structure of DARPA data set is introduced. The application is used to complete the data set with additional features according to an attack analysis. Finally, to show advantages of an extended data set, two genetic methods in the detection of non-content based attacks are tested.

Keywords Adaptative algorithm · Genetic algorithms · Information security · Intrusion detection · Machine learning · TCPIP

30.1 Introduction

Since the beginning of information transmission by means of computer network resources, the security threats have been raised with different approaches, being the most used the Intrusion Detection Systems–IDS [1]. Attacks have been detected with rules obtained by experience or with the use of machine learning algorithms according to data sets acquired from networking scenarios with normal

E. P. Guillén (✉) · J. Rodríguez Parra
Telecomunications Engineering Department, Military University "Nueva Granada",
Cra 11 No. 101-80, Bogotá, Colombia
e-mail: edward.guillen@unimilitar.edu.co

J. Rodríguez Parra
e-mail: gissic@unimilitar.edu.co

R. V. Paéz Mendez
Systems Engineering Department, Javeriana University, Cra 7 # 40-62, Bogotá, Colombia
e-mail: paez-r@javeriana.edu.co

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_30,

traffic and with injected attacks to the same scenario [2–8]. The most widely used data set is DARPA data set developed by MIT [9, 10] with a complete feature collection of attacks, unfortunately some authors have shown weaknesses on the data set either in the lack of important features for detecting some attacks as well as their outdated information [11, 12].

In order to improve detection results, a new set of features have been proposed by some authors, for example with a tool called TSTAT [13]. In the same way, it is possible to find tools supplied by MIT such as LARIAT. The tool simulates traffic generation and allows the aggregations of host and services with the programmed deploy of attacks [14].

Attacks have changed over time, and intruders have found new ways to exploit vulnerabilities without causing detectable variation in the traditional checked features, but other traffic features might be affected. The observed changes in the way the attacks are performed suggest that if new relevant features are included in datasets, it could be possible to improve detection performance.

This chapter shows a software to obtain DARPA compatible features and the possibility of getting additional information in order to construct an updated data set with multiple networking scenarios.

A dataset for detecting non-content based attacks was obtained and two genetic detection methods will be shown in order to analyze the results not only in a simulation scenario but also in the real world. The obtained results enforce the idea of develop hybrid detection systems, according to the characteristics of identified groups of attacks.

In Sect. 30. 2, the obtained database will be explained, as well as a brief introduction to *Spleen* software which was used to obtain DARPA compatible features with new variables. The first version of this work was presented in Guillen et al. [15].

A rule-based Genetic Algorithm–GA attack detection approach will be shown in Sect. 30.3. The weight-based genetic algorithm attack detection is analyzed in Sect. 30.4 to finally show the results in Sect. 30.5 and discuss some conclusions at the end of the chapter.

30.2 Data Base for Attack Analysis

In 1998 a well-known data set for intrusion detection evaluation was developed by the DARPA Intrusion Detection Evaluation Group at MIT Lincoln Laboratory. The data set is composed by 41 statistical, behavioral and status variables collected from network scenarios in presence of diverse attack types [9, 10]. Although the collection of data was finished more than 10 years ago, the data handling standard is useful for analyze new approaches using Machine Learning or Computational Intelligence in order to find or evaluate intrusion detection methods [16]. However, the DARPA data set not only has the outdated problem but also their variables have shown not to be completed for trustable analysis purposes [11, 12]. In order to achieve better results and taking advantage of the MIT previous work, it was developed a software to obtain the results from DARPA data set plus additional variables including PDU content analysis according to the attack to be studied. The software is called *Spleen* and it is available under request at gissic.umng.edu.co website.

30.2.1 Intrusion Detection Database Description

Samples from DARPA that were used for the training and testing process, are composed by two main components, the first one is a collection of features which describes the event and the second one is a class, which informs the type of the event, that is, if it is a normal or an abnormal behavior, all these samples are organized in a CSV file.

Descriptive features of a recorded event can be classified in three groups: discrete, continuous in percentage and continuous count features.

30.2.2 Spleen Data Set Extractor Software

Spleen software is able to collect information from a network interface and gather them with the DARPA data set architecture. The features could be complemented with additional information according to research requirements. Details about possible variables in *Spleen* are not the subject of this chapter but some of them are going to be analyzed for detection purposes, they will be explained in Sect. 30.2.4. The diagram of *Spleen* modules is shown in Fig. 30.1.

30.2.3 Non–Content Based Attacks

In some attacks, information above to layer 4th is sending, in order to take advantage of an application vulnerability, so the payload length of layer fourth is different to zero. These kinds of attacks are usually called “content based attacks” [15].

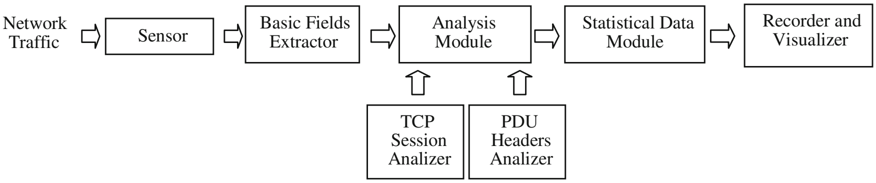


Fig. 30.1 Diagram of Spleen modules

In the other hand, there are some attacks which do not need to send information to the session layer or above, it means that the payload of a possible fourth layer packet is empty; these attacks will be called non-content based attacks and over them a genetic algorithm detection will be shown in order to probe advantages of additional features captured with *Spleen* software.

30.2.4 Variables Selection and Filtering

30.2.4.1 KDD99 General Structure

KDD99 features are divided in 4 categories: Basic, Content based, Statistical based on time and Statistical based on host.

Basic features can be extracted directly from packet headers. They have information about packets source and destination, employed service, payload length and so on.

Content features are in charge of inspecting packet payload, looking for information. Most of this information refers to remote commands sent from machine to machine. There is a lot of ways a command can be sent. A command can be sent by exploiting application vulnerabilities and its format can be altered with self-modifying code. In addition with the no authorized techniques, each protocol has its own way to send and receive its messages. It makes the information gathering difficult to be performed. Even if the message is interpreted in a correct way it is hard to decide if the message by itself is dangerous or not. For all those reasons the KDD99 structure is not efficient on detecting content based attacks [11].

Based on time statistical features refers to the number of events which satisfy some condition in a time slot. The KDD99 sets the time slot to 2 s. Many anomaly detection techniques work in a similar way. They count the number of events in a time slot and based on the results they decide if an intrusion is present or not. But once the duration of the time slot is known by an intruder, the only thing the intruder has to do is to change the interval of his attacks in order to avoid the system.

Based on host statistical features work in a similar way as based on time, but they analyzed the last *n* established connections with a host instead the last connections performed in a time slot. The number *n* is set to 100 in KDD99.

30.2.4.2 Obtained Database

A new database with the data architecture of the DARPA data set was created, and it was completed with some more features in the presence of different attacks with a lab network scenario similar to the one used by MIT [9, 10].

In addition with the feature types presented in the original dataset –KDD99–, a new type of feature is obtained with *Spleen.* The new feature type is called Behavior Change Detector—BCD and it is obtained as follows: during a time slot the events which match some conditions are counted; the process is made again in the next time slot and its results are compared with previous value in last time period. The BCD feature is the difference between the counted events in actual period and in the last time period. It is able to detect abrupt changes in the network traffic behavior.

From the available set, it is necessary to establish the most significant variables in the detection of intrusive events according to the explained focus. In order to reduce the number of examined features and to obtain near real time performance, information gain is calculated. The information gain is based on the concept of entropy. In information theory, entropy is defined as the uncertainty about the nature of an element when it is randomly selected from a dataset with a known probability distribution and it is described by Shannon et al. [17]. The information gain indicates how the knowledge about a feature value can reduce the uncertainty about the traffic type. In order to see this reduction the entropy is calculated twice with different assumptions; in the first case, the feature value is unknown and in the second case it is known. The difference between the results represents the information gain about the feature. Because the Shannon equations were designed for discrete features, all the continuous ones were encoded using the method of equal intervals.

From the 41 obtainable variables, 13 are content based features, it means that such data are not necessary to detect non-content based attacks, i.e., [18] presents a feature relevance analysis over DARPA set to detect attacks and it can be seen that the most important feature in non-content based attacks detection is the status flag of the connection, that is the 4th feature in the data set. The most important possible states of status flag are shown in Table 30.1 [18, 20].

Table 30.1 Status flag possible states

State	Meaning
SF	Normal SYN/FIN completion
REJ	Connection rejected, Initial SYN elicited a RST in reply
S0	State 0: initial SYN seen but no reply
S1	State 1: connection established (SYN's exchanged), nothing further seen
S2	State 2: connection established, initiator has closed their side
S3	State 3: connection established, responder has closed their side
RSTO	Connection reset by the originator
RSTR	Connection reset by the responder
OTH	Other, a state not contemplated here

Another current event to be taken into account is the number of bytes sent to a layer upper than 4th by the originator of the connection –the client, making a two states discretization, when the sent bytes are less than 50 B and when the sent bytes are more than 50 B, because a non-content based attack is characterized by sending few information.

Historical events can help in the detection process because some attacks usually take various steps to be successful. By analyzing the steps, it is possible to know that something is wrong in the network traffic, for example in a portsweep attack, it is possible to find requests made to a not offered service, causing a REJ state connection; this fact by itself does not necessary mean that a portsweep attack is present, instead, it could be possible that someone is trying to get information about the services offered in the network.

Due to the nature of non-content based attacks, the characteristics 4, 25, 26, 27, 28, 29, 34, 39, 40 and 41 were selected to perform the classification [19, 21, 22]. The descriptions of each feature can be found in [11, 18]. As it can be seen, most of the selected features are refereed to statistical information about the TCP connection states.

In order to complete the DARPA data set and by means of statistical analysis, the database was completed using spleen software with the variables illustrated in Table 30.2.

Table 30.2 Some additional features acquired with spleen software

ID	Name	Type	Description
Basic features			
42	Unusual TCP flags	Boolean	True if this connection had received packets with unusual TCP flags configurations
43	Out of sequence packets	Int	The number of packets of this connection that not arrived on time
44	Payload Size average	Int	The size average of the payload length of the packets of this connection, including the retransmissions
Host traffic features (with last 100 connections)			
45	Client count	Int	The number of connections from this client
46	Not completed client rate	Double	% of connections from this client with the states S0 or S1
47	Same host and service	Double	% of connections from this client that have the same host and service
48	Persistent connection rate	Double	% of connections to the current host which were not answered and other attempt to establish them was made
Change behavior detector (difference between two periods)			
49	Host rejected difference	Int	# of connections rejected by the current host
50	Service rejected difference	Int	# of connections to this service that was rejected
51	Unanswered host count	Int	# of connections no answered by this host
52	Unanswered service count	Int	# of connections to the current service that have not been answered yet

30.3 Rule Based GA Attack Detection

An algorithm to detect non- content based attacks was first implemented. With the approach, the main characteristics are focus in headers and behavior from layers third and fourth according to the features explained before. Rule based detection system has the characteristics explained by Holland [23] and Holland et al. [24].

30.3.1 Knowledge Representation

Each line in the data base is represented as follows:

$$S = \{v_1, v_2, \ldots, v_n\}$$

with n number of variables.

Each feature could be represented only by three stages:

$$v_n : \begin{cases} 0 \\ 1 \\ \#\ Dontcare \end{cases}$$

If there is any continuous feature, they are always normalized between 0 and 1, and they are digitalized according to a variability analysis. In that way, v_n could contain more than one bit.

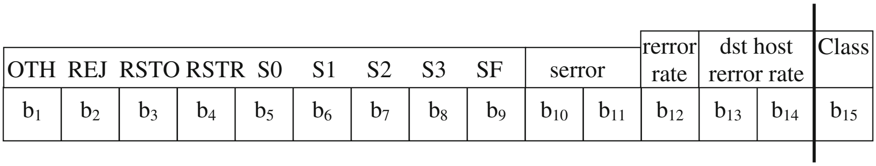
The rules has the {`if` v_1 `and` v_2 `and` v_n `then` class} structure.

The class only has two possible stages "1" or "0" meaning that exists or not an attack. Although the database makes distinction between different attacks such as Teardrope, DoS, Portsweep and so on, all non-content based attacks were merged with just one class.

At the end of each vector, a class of attack or not attack is putting according to database.

30.3.2 The Chromosome

The chromosome has the structure illustrated in Fig. 30.2.



OTH	REJ	RSTO	RSTR	S0	S1	S2	S3	SF	serror		rerror rate	dst host rerror rate		Class
b_1	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}	b_{11}	b_{12}	b_{13}	b_{14}	b_{15}

Fig. 30.2 Initial chromosome structure with just some features

Although the final structure is composed by some more features according to Table 30.2, a chromosome with classical KDD features was implemented in order to compare GA results with additional features obtained with *spleen.*

30.3.3 The Algorithm

The developed approach can be seen in the Fig. 30.3. The fitness function is a measure of effectiveness according to a comparison between each rule and the training data.

A classical mutation operator is used to create a new element based on parents by making a little change in some of its characteristics. To perform this task two parameters are important, the first one is the mutation probability, which defines

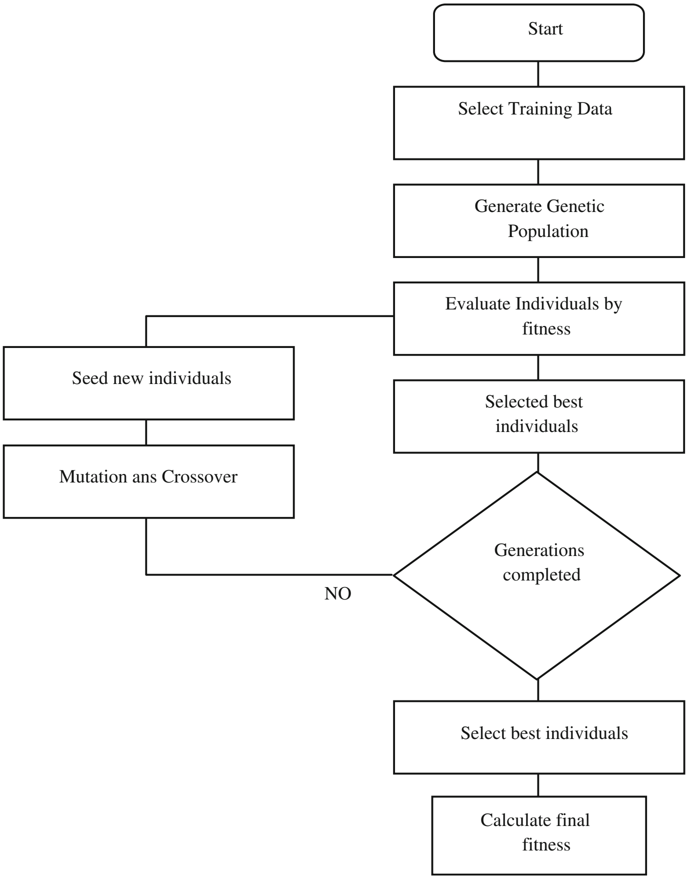


Fig. 30.3 Rule based genetic algorithm

the probability that any of the characteristics could change in new individuals, if it is the case, the second parameter, "deep mutation" defines how many characteristics could change.

The complete database is composed by 150,500 lines. The training data has 4,000 individuals and the validation data was probed with 2,000 individuals randomly selected from the main database. When the best rules were selected, they were tested in an IDS with real time traffic in an attacked network with ethical hacking techniques.

30.4 Weight Based GA Attack Detection

Several features assume special values in presence of some normal or abnormal behaviors; the goal of this approach is to use these values to classify an event. Each feature is going to have an influence level according to its behavior when the event corresponds to a selected class. If the feature takes high values in the events which belong to a selected class, then its influence level is going to be high. Each feature of the event will contribute with a numerical value according to the product of its current value and its influence level. If the sum of all the contributions is bigger than a threshold, then the event is going to be classified as the indicated class. It is important to know that the influence levels can be positive or negative.

Most of the features refer to historical data; however their influence level depends on the current event, for that reason the influence level of each feature will be related with some current condition. Since the most significant feature that represents the current event is the status flag, all features influence levels will have a bonus when this flag is active.

30.4.1 Knowledge Representation

The goal of this genetic algorithm—GA is to find the correct influence level for each feature and to establish what value of status flag is going to give the bonus condition to each feature. With that bonus value is wanted to establish how much the influence level of a feature is going to be incremented in the presence of its bonus status flag, this value is going to be called the bonus value and it is set to 50 % in order to make a significant difference when the bonus status flag is present.

Each influence level is going to be an integer value from −7 to 8 which is going to be represented by 4 bits. The status flag which gives the bonus condition is represented by 4 bits assuming the values shown in Table 30.3.

Another important component of the GA is the threshold value; it is initially set to the number of selected features for the evaluation, allowing each feature to have a relevant role in the detection.

Table 30.3 Possible bonus status flags

Bonus state	Meaning	Bonus state	Meaning
0000	SF	1000	OTH
0001	S3	1001	S0 or S1
0010	S2	1010	RST events
0011	S1	1011	No S0 or S1
0100	S0	1100	No RST events
0101	RSTO	1101	No SF
0110	RSTR	1110	None
0111	REJ	1111	Any

30.4.2 The Chromosome

Each one of selected features, except the status flag of the connection, corresponds to a gene and it is composed by two fields: *influence level* (Weight) and *bonus status* flag (Bonus). The threshold value is the same for all the individuals in order to reduce the GA search space. Chromosome structure for this approach is shown in Fig. 30.4.

30.4.3 The Algorithm

The Weight based GA algorithm is similar to the rule based GA, the main difference is the way the events are evaluated for classification. With the chromosome structure of Fig. 30.4, the weighted sum is evaluated with the Eq. (30.1). The result is then compared with the threshold value. If the threshold is exceeded then the event is labeled according to the class that is wanted to be detected.

$$a = \sum_{i=0}^{n-1} f_i * W_i * (1 + BFS_i * BP) \tag{30.1}$$

where n is the number of selected characteristics, f is the feature value obtained from the training data element, W is the assigned weight to the feature, BFS is the status of the bonus flag assigned for that characteristic "1" or "0", and BP is the bonus value, (BP is the same for all features and elements) Note that the term in parentheses is 1 when the bonus condition –BFS-is 0.

<table>
<tr><td colspan="18">Chromosome</td></tr>
<tr><td colspan="8">Gene 0</td><td>...</td><td colspan="8">Gene (n-1)</td><td>threshold</td></tr>
<tr><td colspan="4">Weight</td><td colspan="4">Bonus</td><td></td><td colspan="4">Weight</td><td colspan="4">Bonus</td><td></td></tr>
<tr><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>b</td><td>Int</td></tr>
</table>

Fig. 30.4 Chromosome structure for weight based GA

Chromosomes															*F* train	*F* test
#	#	0	#	0	0	0	0	0	0	0	#	#	#	1	0,91572	0,97974
#	#	#	#	0	0	0	0	#	0	0	1	#	#	1	0,91069	0,99263
0	#	#	#	#	#	#	0	#	#	0	1	#	#	1	0,91069	1

Fig. 30.5 Example of best rules with some features and the fitness results in training and test process

30.5 Results

The algorithm was tested with an initial population *P* of 50 individuals and with 20 generations *g*. The number of generations and population were tested with different values in order to obtain the best value with the minimal computational cost. Examples of the best obtained rules with the first 14 bits and the class bit with the fitness value *F* in the training and test stages can be seen in Fig. 30.5.

With the data set obtained with *spleen*, the two GA approaches were trained. Every non-content based attack was labeled in just one class. Table 30.4 shows the results of weight based GA with just Darpa data set.

With the same data set but with the additional features the results in Table 30.5 were obtained.

Obtained results were made with next GA parameters: Population of 120, 15 % of mutation probability and a deep of 2.

The same process was performed with the rule based GA, and results appears in Tables 30.6 and 30.7.

Table 30.4 Results of weight based GA only with DARPA

Figure of merit	Training phase	Testing phase
True positives	0.98122	0.9788
True negatives	0.99643	0.98620
False positives	0.00356	0.01379
False negatives	0.01877	0.02111
Fitness	0.98986	0.97985

Table 30.5 Results of weight based GA with the additional data

Figure of merit	Training phase	Testing phase
True positives	0.99530	0.98204
True negatives	1.0	1.0
False positives	0.0	0.0
False negatives	0.00469	0.01795
Fitness	0.99797	0.98443

Table 30.6 Results of rule based GA Using DARPA features

Figure of merit	Training phase	Testing phase
True positives	0.9788	0.97465
True negatives	0.99643	0.97241
False positives	0.00356	0.02534
False negatives	0.02112	0.02758
Fitness	0.98885	0.97435

Table 30.7 Results of rule based GA using the additional data

Figure of merit	Training phase	Testing phase
True positives	0.99530	0.99310
True negatives	1.0	1.0
False positives	0.0	0.0
False negatives	0.00469	0.00689
Fitness	0.99797	0.999084

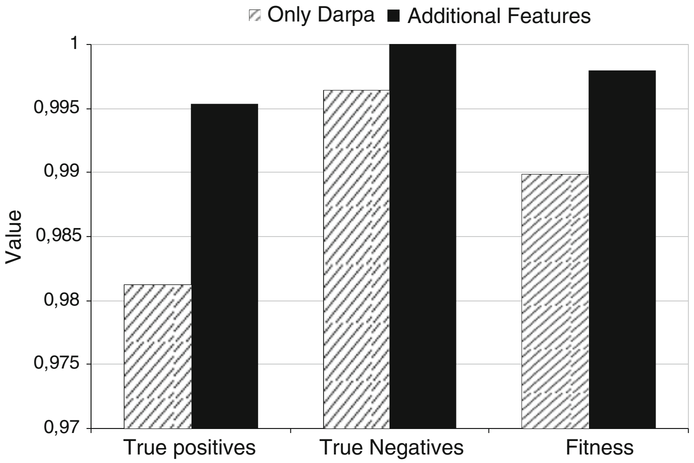


Fig. 30.6 Comparison of results with GA weight based approach with KDD99 classical features and with proposed additional features

In Fig. 30.6, it is shown a comparison of results with GA weight based approach with KDD and with proposed additional features. On the other hand the Fig. 30.7 shows the results obtained by the rule based GA approach.

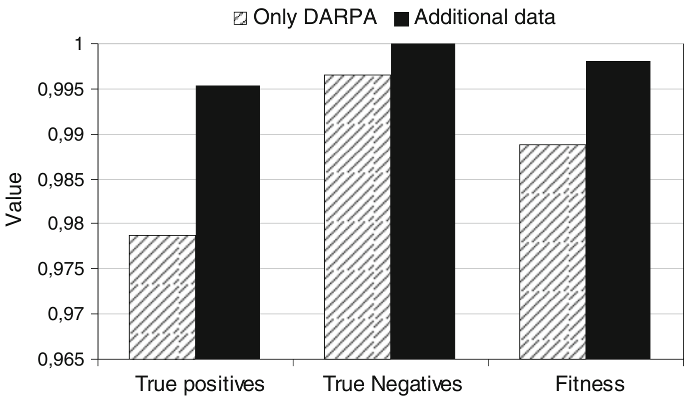


Fig. 30.7 Comparison of results with GA rule based approach with KDD99 classical features and with the addition of proposed features

30.6 Conclusion and Future Work

Spleen provides the possibility to change the way the features are obtained, for example extending the time window or the number of considered connections in statistical measurements. It represents a bigger range of possibilities for research purposes.

The definition of "normal traffic" can differ from network to network; *Spleen* allows the construction of particular datasets which describe the behavior of tested networks. It makes easier the work of IDS schemes to detect abnormal activity.

Future studies about intrusions behavior could possibly give information about what data are useful for their detection, and then modifications to *Spleen* and machine learning algorithms can be made in order to improve detection results.

The new type of feature—behavior change detector- provides a way to detect abrupt changes on network traffic, but it must be implemented carefully according to network traffic distribution. The most important parameter to the behavior change detector variables is the observation time, the longer this period, the more descriptive the measure.

Since intrusions get more sophisticated day after day, *spleen* software could be used to getting new features in order to detect attacks, but it is still necessary to construct a new complete database each time a feature is discovered, that is why it is important to construct a very large data set with as many variables as possible continuously updated.

Acknowledgments This work was possible with the support of Military University Doctoral Support Program, and Javeriana University Doctoral Program.

References

1. Garuba M, Liu C, Fraites D (2008) Intrusion techniques: comparative study of network intrusion detection systems. In: 5th international conference on information technology: new generations, 2008. ITNG 2008, pp 592–598, 7–9 Apr 2008, doi: 10.1109/ITNG.2008.231
2. Ashfaq S, Farooq MU, Karim A (2006) Efficient rule generation for cost-sensitive misuse detection using genetic algorithms. In: 2006 international conference on computational intelligence and security, vol 1, pp 282–285, Nov 2006
3. Shun J, Malki HA (2008) Network intrusion detection system using neural networks. In: 4th international conference on natural computation, 2008, ICNC'08, vol 5, pp 242–246, Oct 2008
4. Devaraju S, Ramakrishnan S (2011) Performance analysis of intrusion detection system using various neural network classifiers. In: 2011 international conference on recent trends in information technology (ICRTIT), pp 1033–1038, June 2011
5. Momenzadeh A, Javadi HHS, Dezfouli MA (2009) Design an efficient system for intrusion detection via evolutionary fuzzy system. In: 11th international conference on computer modelling and simulation, 2009, UKSIM'09, pp 89–94, March 2009
6. Kim DS, Nguyen H-N, Park JS (2005) Genetic algorithm to improve svm based network intrusion detection system. In: 19th international conference on advanced information networking and applications, 2005, AINA 2005, vol 2, pp 155–158, March 2005
7. Ahmed A, Lisitsa A, Dixon C (2011) A misuse-based network intrusion detection system using temporal logic and stream processing. In: 5th international conference on network and system security (NSS), 2011, pp 1–8, Sept 2011
8. Spafford EH, Kumar S A pattern matching model for misuse intrusion detection. Department of computer science
9. MIT Lincoln Laboratory (1999) Darpa intrusion detection data sets
10. Graf I, Haines JW, Kendall KR, McClung D, Weber D, Webster SE, Wyschogrod D, Cunningham RK, Lippmann RP, Fried DJ, Zissman MA (1999) Evaluating intrusion detection systems: the 1998 darpa off-line intrusion detection evaluation. Lincoln Laboratory MIT, 244 Wood Street, Lexington, MA 02173-9108, p 15, 1999
11. Sabhnani M, Serpen G Why machine learning algorithms fail in misuse detection on KDD intrusion detection data set. The University of Toledo
12. Vasudevan AR, Harshini E, Selvakumar S (2011) Ssenet-2011: a network intrusion detection system dataset and its comparison with kdd cup 99 dataset. In: 2nd Asian Himalayas international conference on internet (AH-ICI), pp 1–5, Nov 2011
13. Munafo MM, Mellia M (2008) Tstat measures: Tcp statistics an analysis tool
14. Haines JW, Rossey LM, Lippmann RP, Cunningham RK (2001) Extending the darpa off-line intrusion detection evaluations. In: Proceedings of DARPA Information Survivability Conference Exposition II, 2001, DISCEX'01, vol 1, pp 35–45
15. Guillen E, Rodríguez J, Paez R, Rodríguez A (2012) Detection of non-content based attacks using GA with extended KDD features. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct 2012, San Francisco, USA, pp 30–35
16. Abdullah B, Abd-alghafar I, Salama GI, Abd-alhafez A (2009) Performance evaluation of a genetic algorithm based approach to network intrusion detection system. In: 13th international conference on aerospace sciences and aviation technology, Military Technical College, Kobry Elkobbah, Cairo, Egypt
17. Shannon CE, Weaver W, Blahut RE (1949) The mathematical theory of communication, vol 117. University of Illinois press, Urbana
18. Kayacik HG, Zincir-Heywood AN, Heywood MI (2006) Selecting features for intrusion detection: a feature relevance analysis on KDD 99 intrusion Detection Datasets. Dalhousie University, Faculty of Computer Science, 6050 University Avenue, Halifax, Nova Scotia, 2006

19. Fowdar J, Crockett K, Bandar Z, O'Shea J (2005) On the use of fuzzy trees for solving classification problems with numeric outcomes. In: The 14th IEEE international conference on fuzzy systems, 2005, FUZZ '05, pp 436, 25–25 May 2005
20. Information Sciences Institute University of Southern California. Rfc 793. transmission control protocol. Defense Advanced Research Projects Agency, 1981
21. Hernández-Pereira E, Suárez-Romero JA, Fontenla-Romero O, Alonso-Betanzos A (2009) Conversion methods for symbolic features: a comparison applied to an intrusion detection problem. Expert Syst Appl 36(7):10612–10617
22. Nmap (2012) Port scanning techniques: nmap reference guide, May 2012
23. Holland J (1975) Adaptation in natural and artificial. The University of Michigan Press, Ann Arbor
24. Holland JH et al (2000) What is a learning classifier system? In: Lanzi PL, Stolzmann W, Wilson SW (eds) Learning classifier systems, from foundations to applications. Springer-Verlag, London, pp 3–32

Chapter 31
On the Numerical Solutions of Boundary Value Problems in the Plane for the Electrical Impedance Equation: A Pseudoanalytic Approach for Non-Smooth Domains

Cesar Marco Antonio Robles Gonzalez,
Ariana Guadalupe Bucio Ramirez,
Marco Pedro Ramirez Tachiquin
and Victor Daniel Sanchez Nava

Abstract We study the Electrical Impedance Equation within a special class of domains, whose boundaries posses non-smooth points. The forward Dirichlet boundary value problem is solved bias a novel numerical method, based upon the Pseudoanalytic Function Theory, that does not require additional regularization techniques to fulfill the boundary condition at the non-smooth points.

Keywords Bers · Boundary · Impedance · Non-smooth · Pseudoanalytic · Vekua

Every author equally contributed to the research work.

C. M. A. Robles Gonzalez
ESIME-IPN, Philadelphia, Mexico
e-mail: cesar.robles@lasallistas.org.mx

A. G. Bucio Ramirez (✉)
UPIITA-IPN, Mexico, Mexico
e-mail: ari.bucio@gmail.com

M. P. Ramirez Tachiquin · V. D. Sanchez Nava
The Communications and Digital Signal Processing Group, La Salle University, Philadelphia, Mexico
e-mail: marco.ramirez@lasallistas.org.mx

V. D. Sanchez Nava
e-mail: ddansanchez@gmail.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_31,

31.1 Introduction

The study of the solutions of the forward Dirichlet boundary value problem in the plane, corresponding to the Electrical Impedance equation

$$\text{div}\,(\sigma\ \text{grad}\ u) = 0, \tag{31.1}$$

where $\sigma = \sigma_1(x)\sigma_2(y)$ is the conductivity function, and u is the electric potential, constitutes the base for analyzing the inverse problem, commonly known as Electrical Impedance Tomography.

The discovery of the relation between (31.1) in the plane, and the Vekua equation [9], by Kravchenko [6], and shortly after by Astala and Päivärinta [1], opened a complete new path for constructing numerical solutions of the forward problem corresponding to (31.1), based upon the modern Pseudoanalytic Function Theory [2]. This work is fully dedicated to emphasize a special property belonging to a novel numerical method [7, 8], that achieves to approach solutions of the forward boundary value problem of (31.1), when considering a certain class of bounded domains with non-smooth boundaries, without employing additional regularization methods.

The reader can appreciate that, even there is not a criteria for adequately identify the class of non-smooth boundaries that can be analyzed using this method, it is possible to infer that it will be valid for a wide variety of examples strongly related with physical cases.

31.2 Preliminaries

As it has been properly shown in a variety of works (see e.g. [1, 5]), the two-dimensional case of the Electrical Impedance Eq. (31.1) is fully equivalent to a Vekua equation of the form

$$\partial_{\bar{z}} W - \frac{\partial_{\bar{z}} p}{p}\overline{W} = 0, \tag{31.2}$$

where

$$W = \sqrt{\sigma}(\partial_x u - i\partial_y u),\ \ \partial_{\bar{z}} = \partial_x + i\partial_y, \quad \text{and} \quad p = \sqrt{\sigma_1(x)^{-1}\sigma_2(y)}.$$

Professor Bers [2] showed that the general solution of the Eq. (31.2) can be expressed in terms of the so-called Taylor series in formal powers:

$$W = \sum_{n=0}^{\infty} Z^{(n)}(a_n, z_0; z). \tag{31.3}$$

Here a_n and z_0 are complex constants, $z = x + iy$ and $i^2 = -1$. A detailed description of the analytic construction of $Z^{(n)}(a_n, z_0; z)$ can be found in [2] and [5].

Since in [8] was proved that any arbitrary conductivity function σ, can be considered as the limiting case of a piece-wise separable variables function, at every point within a bounded domain, it is possible to numerically approach the expressions $Z^{(n)}(a_n, z_0; z)$ over a finite set of radii R, considering $z_0 = 0$, according to the expressions:

$$\begin{aligned} Z^{(n+1)}[k] = {} & AF[k] \;\; \mathrm{Re} \sum_{q=1}^{k} \Big(G^*[q] Z^{(n)}[q] + G^*[q+1] Z^{(n)}[q+1] \Big) \Delta z[q] \\ & + AG[k] \;\; \mathrm{Re} \sum_{q=1}^{k} \Big(F^*[q] Z^{(n)}[q] + F^*[q+1] Z^{(n)}[q+1] \Big) \Delta z[q], \end{aligned} \tag{31.4}$$

where

$$\Delta z[q] = x[q+1] - x[q] + i(y[q+1] - y[q]),$$

and q, as well k, represent the points located along every the radius R (see [8]). Finally, A is a real constant that contributes to the numerical stability of the method. As a matter of fact, the expression (31.4) is only a variant of a trapezoidal parametric integration method on the complex plane. A detailed description of this method can be found in [7].

31.3 Analysis of Special Non-Smooth Boundaries

The next paragraphs emphasize a particular property of the numerical method posed in [7], related to the calculations performed for special bounded domains Ω, whose boundaries Γ are non-smooth rectifiable curves, as those suggested in [4]. In general, when studying these kind of boundaries, additional regularization techniques are required at the non-smoothness points, for adequately approaching solutions of boundary value problems of (31.1). Nevertheless, the examples displayed further, show two cases where not any regularization method is required for solving the forward Dirichlet boundary value problem, even when the conductivity within the bounded domain is composed by non-smooth geometrical figures.

These are not unique examples, but it has not been established a full criteria to distinguish the class of non-smooth boundaries, that are susceptible to this particular analysis. Still, the very fact that the numerical method can approach solutions for the forward Dirichlet boundary value problem of (31.1), considering non-smooth boundaries, without employing additional regularization techniques (first noticed in [8]), is an interesting topic to analyze.

We shall begin the discussion studying a non-smooth bounded domain, where the conductivity function possesses an exact mathematical representation.

31.3.1 A Case When the Conductivity Possesses a Lorentzian Form

The Fig. 31.1 displays a domain consisting into a circle with radius $R = 1$, within the closed interval $x \in [-1, \cos 0.1\pi]$; and a triangle whose sides are traced by the segments of the lines $f_0(x) = \cos 0.1\pi$, $f_1(x) = \alpha x + \beta$ and $f_2(x) = -\alpha x + \beta$. More precisely, for the case plotted in Fig. 31.1, we have that $\alpha = 0.5629$ and $\beta = 0.8443$. Yet, hereafter it will result more convenient to employ the distance between the center of the unitary circle, and the intersection of the lines $f_1(x)$ and $f_2(x)$, denoted as b, for illustrating the behavior of the numerical method.

The calculations were performed considering 200 radii, equally distributed among the angle interval $\theta \in [0, 2\pi]$; and 201 points equally distributed along every radius, being the first point located at the center of the circle, and the last at the intersection of the radius with the boundary Γ; and a maximum of $2N + 2 = 61$ formal powers $Z^{(n)}(a_n, z_0; z)$.

It is necessary to remark that not any significant variation of the convergence was observed when increasing the number of points per radius, whereas the increment of the total number of radii or the maximum number of formal powers, does provoke a significant diminution in the accuracy of the method. The explanation of this particular behavior is still a topic of research.

The imposed boundary condition will be

$$u|_{\Gamma} = \frac{1}{3}\left(x^3 + y^3\right) + 0.5(x + y), \tag{31.5}$$

which is indeed an exact solution of the Eq. (31.1), for the case when the conductivity σ possesses a Lorentzial form:

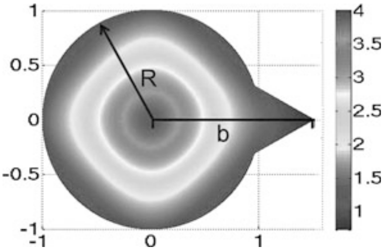


Fig. 31.1 Domain with three non-smooth points at the boundary, containing a Lorentzian conductivity $\sigma = (x^2 + 0.5)^{-1}(y^2 + 0.5)^{-1}$. One radius was forced to intersect the boundary Γ in every non-smooth point, thus the non-smoothness is considered in the numerical analysis

Table 31.1 Relations among the number M of base functions for approaching the solution $u|_{app}$, the parameter b and the error $\mathscr{E}$, corresponding to the domain illustrated in Fig. 31.1

M	b	$\mathscr{E}$
61	1.0	1.3187×10^{-3}
41	1.0	2.6002×10^{-3}
21	1.0	9.3481×10^{-3}
61	1.5	2.2615×10^{-3}
41	1.5	3.5889×10^{-3}
21	1.5	1.2248×10^{-2}
61	2.0	1.6529×10^{-2}
41	2.0	1.4621×10^{-2}
21	2.0	3.2788×10^{-2}

$$\sigma = \left(x^2 + 0.5\right)^{-1}\left(y^2 + 0.5\right)^{-1}. \tag{31.6}$$

Notice every point (x, y) on which (31.5) is valued, belongs to the boundary Γ. The criteria of convergence, considered in this work, will be the error $\mathscr{E}$ obtained bias the Lebesgue norm:

$$\mathscr{E} = \left(\int_{\Gamma}\left(u|_{\Gamma} - u|_{app}\right)^2 dl\right)^{\frac{1}{2}},$$

where $u|_{app}$ represents the approached solution, constructed employing the numerical formal powers obtained by the formulas (31.4). We refer the reader to the works [3] and [8] for a complete and detailed explanation of the algorithm that allows the construction of $M = 2N + 1$ orthonormal base functions at the boundary Γ, whose linear combination reaches $u|_{app}$, once $2N + 2$ numerical formal powers $Z^{(n)}(a_n, z_0; z)$ have been approached.

The Table 31.1 displays the behavior of the error $\mathscr{E}$ when changing the number M of base functions, and the magnitude of the parameter b introduced above (plotted in Fig. 31.1).

31.3.2 A Case When the Conductivity is Provided by Geometrical Figures

As a matter of fact, this case is more representative for engineering applications, because the conductivity of a physical experiment is rarely expressed by an exact mathematical formula. Moreover, the classical numerical methods for solving boundary value problems of elliptic partial differential equations, could require regularization techniques of considerable complexity, for analyzing the geometrical conductivity displayed in Fig. 31.2, due to the presence of non-smoothness into the figure within the bounded domain.

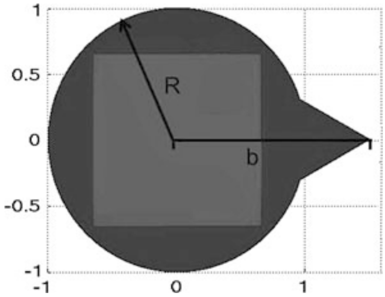


Fig. 31.2 Domain with non-smooth boundary, containing a conductivity σ provided by geometrical figures. Notice the figure within the domain also possesses a non-smooth perimeter. Once more, every non-smoothness point is crossed by one radius, without employing additional regularization methods for solving the boundary value problem

Table 31.2 Relations among the number M of base functions for approaching the solution $u|_{app}$, the parameter b and the error $\mathscr{E}$, corresponding to the domain illustrated in Fig. 31.2

M	b	$\mathscr{E}$
61	1.0	1.8931×10^{-2}
41	1.0	3.3239×10^{-2}
21	1.0	7.8480×10^{-2}
61	1.5	2.5619×10^{-2}
41	1.5	4.3609×10^{-2}
10	1.5	9.1848×10^{-2}
61	2.0	1.0997×10^{-1}
41	2.0	6.1221×10^{-2}
21	2.0	1.2170×10^{-1}

Once more, we fixed the number of radii in 200, the number of points per radius in 201, and the maximum number of formal powers in $2N + 2 = 62$; noticing that the increment of the number of radii or the maximum number of formal powers, as in the previous example, also provokes a diminution of the convergence.

We shall remark that for every point of non-smoothness, at the boundary Γ as well at the figure inside the domain, one radius was forced to cross through. Thus every non-smoothness is effectively taken into account. The Table 31.2 contains the information about the error $\mathscr{E}$ when introducing changes in the total number of base functions M and into the parameter b, emphasizing that not any additional regularization technique was employed to improve the convergence of the method.

Acknowledgments The authors would like to acknowledge the support of CONACyT projects 106722 and 81599; A. G. Bucio Ramirez thanks to UPIITA-IPN and CONACyT; M.P. Ramirez Tachiquin acknowledges the support of HILMA S.A. de C.V.; C. M. A. Robles-Gonzalez and V. D. Sanchez Nava would like to thank La Salle University for the research stay and to CONACyT.

References

1. Astala K, Päivärinta L (2006) Calderon's inverse conductivity problem in the plane. Ann Math 163:265–299
2. Bers L (1953) Theory of pseudoanalytic functions. New York University, New York, IMM
3. Bucio AR, Castillo-Perez R, Ramirez MPT (2011) On the numerical construction of formal powers and their application to the electrical impedance equation. In: 8th international conference on electrical engineering, computing science and automatic control. IEEE Catalog Number: CFP11827-ART, ISBN:978-1-4577-1013-1, pp 769-774
4. Castillo-Perez R, Kravchenko V, Resendiz VR (2011) Solution of boundary value and eigenvalue problems for second order elliptic operators in the plane using pseudoanalytic formal powers. Math Methods Appl Sci 34(4):455–468
5. Kravchenko VV (2009) Applied pseudoanalytic function theory. Series: frontiers in mathematics, ISBN 978-3-0346-0003-3
6. Kravchenko VV (2005) On the relation of pseudoanalytic function theory to the two-dimensional stationary Schrödinger equation and Taylor series in formal powers for its solutions. J Phys A: Math Gen 38(18):3947–3964
7. Robles G CMA, Bucio R A, Ramirez T MP (2012) An optimized numerical method for solving the two-dimensional impedance equation.In: Proceedings of The world congress on engineering and computer science (2012) WCECS 2012, 24–26 October 2012. Lecture notes in engineering and computer science, USA, San Francisco, pp 116–121
8. Ramirez T MP, Hernandez-Becerril RA, Robles G MC (2012) Study of the numerical solutions for the electrical impedance equation in the plane: a pseudoanalytic approach of the forward dirichlet boundary value problem. Math Methods Appl Sci (submited for publication). Available in electronic format at http://ArXiv.com
9. Vekua IN (1962) Generalized analytic functions, international series of monographs on pure and applied mathematics. Pergamon Press, London

Chapter 32
An Ontology-Based Methodology for Building and Matching Researchers' Profiles

Nawarat Kamsiang and Twittie Senivongse

Abstract To support potential research collaboration, we present an ontology-based methodology for identifying common research interest among researchers. The methodology uses an ontology building algorithm to build researchers' ontological profiles from publication keywords, and then an ontology matching algorithm is used to identify common research areas and degree of matching between research profiles. Our ontology matching also considers depth weights, i.e., the depth of the ontological terms within the two profiles that are matched. The idea is the terms that are located near the bottom of the ontologies should indicate specialization of researchers, and hence attention should be paid more to matching of such terms than to matching of the terms that are closer to the top of the ontologies. We present an experiment to match profiles of researchers in the same field, close fields, and different fields, and report the performance of the methodology and an evaluation using an ontology matching benchmark. The methodology is considered useful as it can quantify similarity of research interests and give practical matching results.

Keywords Neighbor search algorithm · Ontology building · Ontology matching · Profile matching · Research expertise · WordNet

N. Kamsiang
Office of Computer Service, Sripatum University Chonburi Campus,
Chonburi 20000, Thailand
e-mail: nawarat.km@east.spu.ac.th

T. Senivongse (✉)
Department of Computer Engineering, Chulalongkorn University,
Bangkok 10330, Thailand
e-mail: twittie.s@chula.ac.th

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_32,

32.1 Introduction

Identifying common research interest is a challenging task for promoting research collaboration. Researchers seek collaboration with one another for sharing ideas and resources, complementing one's expertise with others', as well as increasing visibility of the research work and the researchers themselves. Collaboration between researchers in the same field can specifically strengthen the work within the field while collaboration between different fields may lead to useful innovative work with wider application across fields.

The basis for identifying shared research interest is analyzing researchers' expertise and finding correspondence between research areas. Primarily, association between researchers can be drawn using bibliographic data of their publications [1]. Researchers who, for example, co-author publications, cite similar publications, or use similar keywords in publications can be identified as sharing common interest. Another approach taken by large number of related work is gathering researchers' information from electronic sources, e.g., online libraries, Web sites, blogs, and project documents, to build research profiles and mine researchers' expertise.

This chapter presents a methodology for identifying common interest among researchers. As opposed to a thorough analysis on bibliographic and Web-based information of the researchers, the methodology is lightweight since the basis of its analysis is merely a set of publication keywords. The motivations are that, we would like to explore a simple way to identify potential research collaboration, and the collaboration should not be limited to researchers within the same professional circle. In other words, the methodology should be able to identify if any two researchers have potential common interest even when they are in different networks of expertise. Our methodology uses an ontology building algorithm to build researchers' ontological profiles from keywords of previous publications, and then an ontology matching algorithm is used to determine the degree of matching between the profiles and identify matched ontological terms as shared areas of interest. Here the ontology matching algorithm takes into account the depth weights of the matched terms. That is, it is interested in matching of the terms that are located near the bottom of the ontological profiles since they are specialization of the researchers, and should represent common interest better than matched terms that are near the top of the ontological profiles. This chapter is a revised and extended version of the work in [2, 3]. In particular, the chapter revisits the methodology and presents a further experiment to match profiles of researchers in the same field, close fields, and different fields. In addition, the performance of profile matching is evaluated using the OAEI 2012 benchmark [4].

Section 32.2 discusses research work related to this chapter. Section 32.3 gives the detail of the methodology through building and matching the profiles of two researchers. The evaluation of the methodology is presented in Sect. 32.4 followed by a conclusion in Sect. 32.5.

32.2 Related Work

Information from electronic sources, such as bibliographic databases, Web sites, and discussion forums, has been used widely for analyzing expertise and connection between people. A well-known search and mining tool called ArnetMiner [5] can provide search services including researcher profile search, expert finding, active researchers for conferences, and researcher ranking. To build researcher social network, it extracts researchers' information from the Web to create semantic-based profiles for the researchers and has their bibliographic data from several digital libraries integrated with the profiles. Zhang et al. [6] analyze asker-helper interaction in the Java Forum threads by considering the number of replies any user has posted to help others and whom the user has helped. The analysis uses network-based ranking algorithms, including PageRank and HITS, to identify users with expertise. Punnarut and Sriharee [7] build semantic-based researcher profiles based on ACM computing classification and compute expertise scores, find researchers who share expertise, as well as rank them. Trigo [8] finds researchers with similar interest by using the DBLP bibliographic database and research Web pages as the sources for extracting researcher information, and applying text mining techniques to discover relations between them. Yang et al. [9] construct a social network of researchers by analyzing four types of data, i.e., publication keywords, personal interest, themes of the conferences where papers are published, and co-authorship. An interesting finding is that publication keywords can represent research interest better than co-authorship data. Motivated by the related work, we explore another approach to determining common research interest by representing researchers' profiles as ontologies, which are built upon publication keywords. Then we compare the profiles using an ontology matching algorithm.

32.3 Methodology

Given two researchers, our methodology determines how much their interests match as well as the research areas that they share. The methodology comprises two steps: building ontological research profiles and matching the profiles. Here we revisit the methodology through a case of two researchers in the same field.

32.3.1 Building Ontological Research Profiles

A researcher's profile is built upon publication keywords. A researcher's keywords under a particular subject area during a certain period are taken from ISI Web of Knowledge database [10]. (Note that the subject areas referred to in this paper are by ISI categorization.) As an example, we select two researchers in the same Computer Science subject area, i.e., Kijsirikul B and Ratanamahatana CA. We

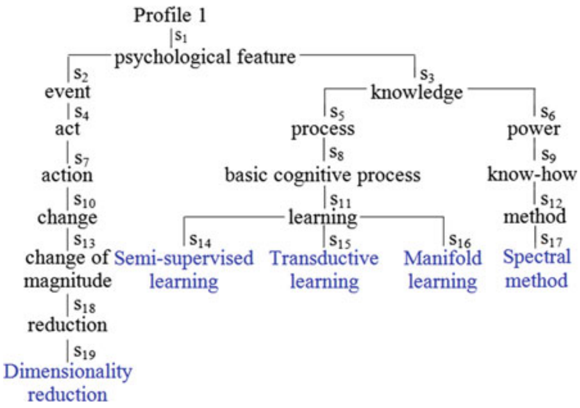


Fig. 32.1 Kijsirikul B's ontological profile

take top five keywords of their publications during the year 2002–2011 as starting terms for building their ontological profiles as depicted in Figs. 32.1 and 32.2. In this section, we describe how to build such profiles. Note that the five starting terms of each researcher are at the bottom of the corresponding profile.

We adopt the Obtaining Domain Ontology (ODO) algorithm proposed by An et al. [11] which can automatically derive a domain-specific ontology from items of information (i.e., keywords in this case). Starting with each keyword, we repeatedly find terms and hypernym (i.e., parent) relation from WordNet [12] to build an ontology fragment as a Directed Acyclic Graph. Since a term may have several hypernyms, for simplicity, we select one with the maximum tag count which denotes that the hypernym of a particular sense (or meaning) is most frequently used and tagged in various semantic concordance texts. For example, in Fig. 32.2, *clustering* has *agglomeration* as its hypernym, and *agglomeration* has *collection* as its hypernym, and so on. If the term is not found in WordNet but is a noun phrase consisting of a head noun and modifier(s), we generalize the term by removing one modifier at a time to look up in WordNet. If found, that generalized form becomes

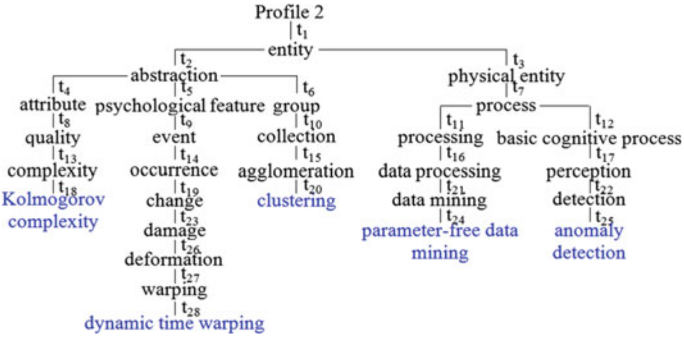


Fig. 32.2 Ratanamahatana CA's ontological profile

the hypernym. Otherwise, the researcher's subject area is used as the hypernym. For example, in Fig. 32.1, the noun phrase *Semi-supervised learning* is not found in WordNet, so we use the head noun *learning* as its hypernym. After we obtain all ontology fragments, we then join together identical terms in different fragments to form an ontology representing a profile of research interest. For example, in Fig. 32.1, five ontology fragments based on five starting terms are joined at the terms *learning*, *knowledge*, and *psychological feature* respectively.

32.3.2 Matching Ontological Research Profiles

After we obtain the ontological profiles of any two researchers, we compare their profiles using ontology matching. The basis of our ontology matching is the algorithm called *Multi-level Matching Algorithm with the neighbor search algorithm* (*MLMA+*) proposed by Alasoud et al. [13] as shown in Fig. 32.3. MLMA+ consists of initialization, neighbor search, and evaluation phases. The detail of these phases is presented in Sects. 32.3.2.1–32.3.2.3. We then replace the initialization phase of MLMA+ with a modification in Sect. 32.3.2.4 for our *MLMA+ with depth weights algorithm.*

32.3.2.1 Initialization Phase

To match an ontology *S* with another ontology *T*, preliminary matching techniques are applied to determine similarity between terms in the two ontologies. The

Fig. 32.3 MLMA+ algorithm [13]

```
Algorithm Match (S, T)
begin
/* Initialization phase
    $K \leftarrow 0$ ;
    $St_0 \leftarrow$ preliminary_matching_techniques (S, T) ;
    $St_f \leftarrow St_0$ ;
/* Neighbor Search phase
    $St \leftarrow$ All_Neighbors ($St_n$) ;
    While (K++ < Max_iteration) do
/* Evaluation phase
        If score ($St_n$) > score ($St_f$) then
            $St_f \leftarrow St_n$ ;
        end if
        Pick the next neighbor $St_n \in St$;
        $St \leftarrow St - St_n$ ;
        If $St = \varnothing$ then return $St_f$ ;
    end
    Return $St_f$ ;
end
```

$$L=\begin{array}{c|cccccccc|}
 & t_1 & t_2 & t_3 & t_4 \dots & t_{14} \dots & t_{23} \dots & t_{25} & \dots \\
s_1 & 0.359 & 0.454 & 0.324 & 0.402 & 0.428 & 0.36 & 0.027 & \dots \\
s_2 & 0.386 & 0.421 & 0.372 & 0.389 & 0.51 & 0.456 & 0.034 & \dots \\
s_3 & 0.341 & 0.417 & 0.286 & 0.413 & 0.419 & 0.37 & 0.039 & \dots \\
\vdots & & & & & & & & \\
s_{10} & 0.405 & 0.446 & 0.339 & 0.419 & 0.524 & 0.567 & 0.032 & .. \\
s_{11} & 0.272 & 0.403 & 0.221 & 0.409 & 0.341 & 0.313 & 0.034 & \dots \\
\vdots & & & & & & & & \\
s_{16} & 0.032 & 0.034 & 0.036 & 0.032 & 0.032 & 0.032 & 0.039 & \dots \\
s_{17} & 0.272 & 0.453 & 0.311 & 0.281 & 0.036 & 0.034 & 0.034 & \dots \\
\vdots & & & & & & & & \\
s_{19} & 0.025 & 0.03 & 0.03 & 0.025 & 0.024 & 0.023 & 0.039 & \dots
\end{array}
\qquad
\mathrm{Map}_{0-1}=\begin{array}{c|cccccccc|}
 & t_1 & t_2 & t_3 & t_4 \dots & t_{14} & t_{23} & t_{25} & \dots \\
s_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
s_2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots \\
s_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\vdots & & & & & & & & \\
s_{10} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & \dots \\
s_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\vdots & & & & & & & & \\
s_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
s_{17} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\
\vdots & & & & & & & & \\
s_{19} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots
\end{array}$$

Fig. 32.4 L and initial Map$_{0-1}$ for Kijsirikul B and Ratanamahatana CA based on MLMA+

matching techniques used here are name matching using Levenshtein distance, and linguistic matching using WordNet. Levenshtein distance determines the minimal number of insertions, deletions, and substitutions to make two strings equal [14]. For linguistic matching, we use Wu and Palmer's measure provided in the Perl module in WordNet::Similarity package [15] to determine semantic similarity between any two terms. For example, to match Kijsirikul B's ontology S which comprises n terms with Ratanamahatana CA's ontology T comprising m terms, a similarity matrix $L(i, j)$ of the size $n \times m$ is computed. This matrix includes values called *similarity coefficients*, ranging between [0,1] and denoting the degree of similarity between the terms s_i in S and t_j in T. A similarity coefficient is computed as an average of Levenshtein distance and linguistic similarity of the two terms. The similarity matrix L for Kijsirikul B and Ratanamahatana CA is shown in Fig. 32.4. The similarity coefficient of the terms *change* (s_{10}) and *damage* (t_{23}) is 0.567; it is an average of Levenshtein distance (0.2) and linguistic similarity (0.933) of the two terms.

Then, a user-defined threshold *th* is applied to the matrix L to create a binary matrix Map_{0-1}. The similarity coefficient that is less than the threshold becomes 0 in Map_{0-1}, otherwise it is 1. That is, the threshold determines which pairs of terms are considered similar or matched by the user. Figure 32.4 also shows Map_{0-1} for Kijsirikul B and Ratanamahatana CA with $th = 0.5$. This Map_{0-1} becomes the initial state St_0 for the neighbor search algorithm.

32.3.2.2 Neighbor Search Phase

Given the initial state St_0, we search in its neighborhood. Each neighbor St_n is computed by toggling a bit of St_0, so the total number of neighbor states is $n \times m$. An example of a neighbor state is in Fig. 32.5.

32.3.2.3 Evaluation Phase

The initial state and all neighbor states are evaluated using the matching score function v (32.1) [13]:

Fig. 32.5 One of the neighbor states of the initial Map$_{0-1}$ in Fig. 32.4

$$\mathrm{Map}_{0-1}=\begin{array}{c|cccccccc|}
 & t_1 & t_2 & t_3 & t_4 .. & t_{14} . & t_{23} . & t_{25} & \ldots \\
s_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
s_2 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \ldots \\
s_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
\vdots & & & & & & & & \\
s_{10} & 0 & 0 & 0 & 0 & 1 & 1 & 0 & \ldots \\
s_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
\vdots & & & & & & & & \\
s_{16} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \ldots \\
s_{17} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
\vdots & & & & & & & & \\
s_{19} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots
\end{array}$$

$$v = (Map_{0-1} \cdot L)/k = \sum_{i=1}^{n}\sum_{j=1}^{m} Map_{0-1}(i,j) \cdot L(i,j) \Big/ \sum_{i=1}^{n}\sum_{j=1}^{m} Map_{0-1}(i,j);\ v \geq th \tag{32.1}$$

where k is the number of matched pairs and Map_{0-1} is St_n. The state St_n with the maximum score (i.e., St_f of Fig. 32.3) is the answer to the matching. That is, any matched pairs (s_i, t_j) in St_f with bit 1 become common research areas, and its matching score represents the degree of matching between S and T.

32.3.2.4 Modification with Depth Weights

Inspired by the concept of semantic distance between ontological terms [16], we modify the initialization phase of MLMA+ by adding the concept of depth weights and introducing our *MLMA+ with depth weights* algorithm. A depth weight of a pair of matched terms is determined by the distance of the terms from the root of their ontologies. We are interested particularly in matching of the terms that are located near the bottom of the ontological profiles, since they are specialization of the researchers, and should represent common interest better than matched terms that are near the top of the ontological profiles. In Fig. 32.4, consider $s_2 = event$ and $t_{14} = occurrence$. The two terms have similarity coefficient = 0.51. They are relatively more generalized terms in the profiles compared to the pair $s_{10} = change$ and $t_{23} = damage$ with similarity coefficient = 0.567. But both pairs are equally considered as matched areas of interest. We are in favor of the matched pairs that are more specialized and are motivated to decrease the similarity coefficient of the generalized matched pairs by using a depth weight function w (32.2):

$$w_{ij} = \big(rdepth(s_i) + rdepth(t_j)\big)/2;\, w_{ij}\,\text{is in}\ (0, 1] \tag{32.2}$$

where $rdepth(t)$ = relative distance of the term t from the root of its ontology

$$= \frac{\text{depth of the term } t \text{ in its ontology}}{\text{height of ontology}}.$$

$$L = \begin{array}{c|cccccccc|} & t_1 & t_2 & t_3 & t_4 \ldots & t_{14} \ldots & t_{23} \ldots & t_{25} & \ldots \\ \hline s_1 & 0.04 & 0.074 & 0.053 & 0.085 & 0.134 & 0.149 & 0.011 & \ldots \\ s_2 & 0.068 & 0.095 & 0.084 & 0.107 & 0.191 & 0.217 & 0.016 & \ldots \\ s_3 & 0.06 & 0.094 & 0.064 & 0.114 & 0.157 & 0.176 & 0.019 & \ldots \\ \vdots & & & & & & & & \\ s_{10} & 0.147 & 0.184 & 0.14 & 0.194 & 0.295 & 0.376 & 0.021 & .. \\ s_{11} & 0.099 & 0.166 & 0.091 & 0.189 & 0.192 & 0.207 & 0.023 & \ldots \\ \vdots & & & & & & & & \\ s_{16} & 0.014 & 0.016 & 0.017 & 0.017 & 0.02 & 0.023 & 0.028 & \ldots \\ s_{17} & 0.133 & 0.243 & 0.167 & 0.165 & 0.023 & 0.025 & 0.025 & \ldots \\ \vdots & & & & & & & & \\ s_{19} & 0.014 & 0.018 & 0.018 & 0.016 & 0.018 & 0.02 & 0.033 & \ldots \\ \hline \end{array} \quad \mathrm{Map}_{0-1} = \begin{array}{c|cccccccc|} & t_1 & t_2 & t_3 & t_4 \ldots & t_{14} \ldots & t_{23} \ldots & t_{25} & \ldots \\ \hline s_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ s_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ s_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & \\ s_{10} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \ldots \\ s_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & \\ s_{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ s_{17} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \vdots & & & & & & & & \\ s_{19} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ldots \\ \hline \end{array}$$

Fig. 32.6 L and initial $\mathrm{Map}_{0\text{–}1}$ for Kijsirikul B and Ratanamahatana CA based on MLMA+ with depth weights

The original similarity coefficient between s_i and t_j will be multiplied by their depth weight to get a *weighted similarity coefficient*. Therefore the similarity matrix *L(i, j)* would change to include weighted similarity coefficients of the terms s_i and t_j instead.

For $s_2 = \textit{event}$ and $t_{14} = \textit{occurrence}$ in Figs. 32.1 and 32.2, $\textit{rdepth}(s_2) = 2/8$ *and* $\textit{rdepth}(t_{14}) = 5/10$. Their depth weight w would be 0.375 and hence their weighted similarity coefficient would change from 0.51 to 0.191 (0.375 × 0.51). But for $s_{10} = \textit{change}$ and $t_{23} = \textit{damage}$, $\textit{rdepth}(s_{10}) = 5/8$ *and* $\textit{rdepth}(t_{23}) = 7/10$. Their depth weight w would be 0.663 and hence their weighted similarity coefficient would change from 0.567 to 0.376 (0.663 × 0.567). It is seen that the more generalized the matched terms, the more they are "penalized" by the depth weight. Any matched terms that are both the terminal node of the ontology would not be penalized (i.e., $w = 1$). Figure 32.6 shows the new similarity matrix L with weighted similarity coefficients, and the new initial $\textit{Map}_{0\text{–}1}$ for Kijsirikul B and Ratanamahatana CA where $th = 0.3$. Note that for the pair $s_2 = \textit{event}$ and $t_{14} = \textit{occurrence}$ and the pair $s_{10} = \textit{change}$ and $t_{14} = \textit{occurrence}$, they are considered matched in Fig. 32.4 but are relatively too generalized and considered unmatched in Fig. 32.6. For $s_{10} = \textit{change}$ and $t_{23} = \textit{damage}$, they survive the penalty and are considered matched in both figures.

32.4 Evaluation

The evaluation of the methodology is supported by a profile building and matching tool that we have developed in PHP. We present a comparison of matching results produced by the original MLMA+ and the modified MLMA+ with depth weights together with an evaluation of their performance.

32.4.1 Comparison of Matching Results

We compare the profile of Kijsirikul B in Computer Science subject area with a researcher in the same field (i.e., Ratanamahatana CA in Computer Science), a researcher in a close field (i.e., Sudsang A in Robotics whose profile is in Fig. 32.7), and a researcher in a different field (i.e., Ruppin E in Biochemistry & Molecular Biology whose profile is in Fig. 32.8).

Table 32.1 lists the matching scores and matched pairs (s_i, t_j) that represent the shared interests between Kijsirikul B and other researchers. MLMA+ gives a big list of matched pairs which include those very generalized terms near the top of the profiles. Depth weights, on the other hand, lessen the effect of similarity coefficients and hence lower the matching score. As a result, they filter out some generalized matched pairs, giving a more concise list of shared interests which should be more practical for further use.

Table 32.1 also shows that, Kijsirikul B's profile is closest to Ratanamahatana CA's in both algorithms. But it is slightly closer to Sudsang A's when using MLMA+ with depth weights. The matching scores and matched results can vary if the starting keywords (and hence the terms in the profile) as well as the user-defined threshold *th* change. If *th* is too high, the matching score would be very high because most of the matched pairs would be identical terms that are located near the root of the ontologies (e.g., (*psychological feature*, *psychological feature*)). We see that discovering only identical matched pairs are not very interesting given that the benefit of using WordNet and linguistic similarity between non-identical terms would not be present in the matching result. On the contrary, if *th* is too low, there would be proliferation of matched pairs because, even a matched pair is penalized by its depth weight, its weighted similarity coefficient remains greater than the low *th*. The values *th* that we use for the data set in the experiment trades off these two aspects; it is the highest threshold by which the matching result contains both the identical and non-identical matched pairs.

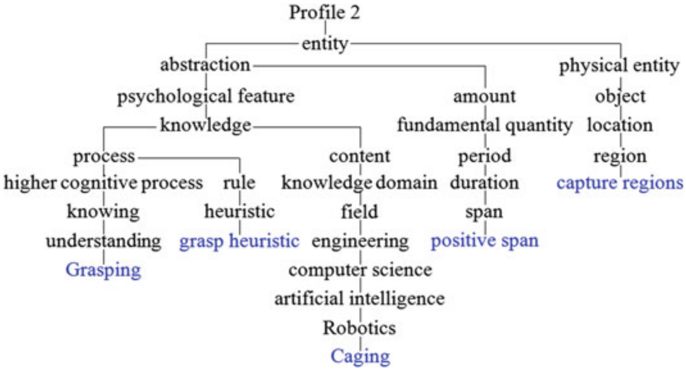


Fig. 32.7 Sudsang A's ontological profile

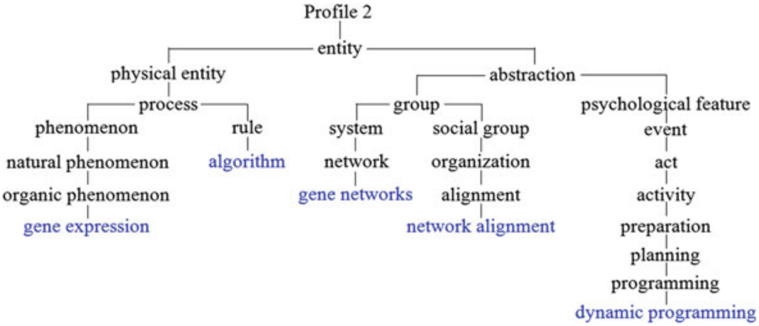


Fig. 32.8 Ruppin E's ontological profile

32.4.2 Performance

The complexity of the ODO algorithm for building an ontology S depends on the number of terms in S and the size of the search space when joining any identical terms in S into single nodes, i.e., $O(\binom{n}{2})$, where the number of ontology terms $n =$ number of starting keywords x depth of S, given that, in the worst case, all starting keywords are of the same depth.

On ontology matching, the complexity of MLMA+ and MLMA+ with depth weights depends on the size of the search space when matching two ontologies S and T, i.e., $O((n \times m)^2)$ when n and m are the size of S and T respectively.

In addition, we evaluate ontology matching using the OAEI 2012 benchmark test sample suite [4]. We use the test sets in the bibliographic domain, each comprising a test ontology in OWL language and a reference alignment. Each test ontology is a modification to the reference ontology #101 and is to be aligned with the reference ontology. Each reference alignment lists expected alignments. So in the test set #101, the reference ontology is matched to itself, and in the test set #n, the test ontology #n is matched to the reference ontology. The quality indicators we use are precision (32.3), recall (32.4), and F-measure (32.5).

$$\text{Precision} = \frac{\text{no. of expected alignments found as matched by algo.}}{\text{no. of matched pairs found by algo.}} \tag{32.3}$$

$$\text{Recall} = \frac{\text{no. of expected alignments found as matched by algo.}}{\text{no. of expected alignments}} \tag{32.4}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{32.5}$$

Table 32.2 shows the evaluation results with $th = 0.5$. We group the test sets into four groups. Test set #101–104 contain test ontologies that are more generalized or restricted than the reference ontology by removing or replacing OWL

Table 32.1 Matching results for Kijsirikul B and other researchers

Kijsirikul B (computer science) vs.	MLMA+ (*th* = 0.5)	MLMA+ with depth weights (*th* = 0.3)
Ratanamahatana CA (computer science)	Score = 0.627 (25 pairs): (psychological feature, psychological feature), (event, event), (event, occurrence), (event, change), (knowledge, process), (power, process), (power, event), (power, quality), (process, process), (process, processing), (act, process), (act, event), (act, change), (action, process), (action, change), (action, detection), (basic cognitive process, basic cognitive process), (change, event), (change, occurrence), (change, change), (change, damage), (change, deformation), (change of magnitude, change), (reduction, change), (knowledge, perception)	Score = 0.385 (14 pairs): (event, event), (process, process), (basic cognitive process, basic cognitive process), (change, change), (change, damage), (change, deformation), (change, warping), (change of magnitude, change), (change of magnitude, damage), (change of magnitude, deformation), (reduction, change), (reduction, detection), (reduction, damage), (reduction, deformation)
Sudsang A (robotics)	Score = 0.581 (24 pairs): (psychological feature, psychological feature), (event, field), (knowledge, knowledge), (knowledge, content), (knowledge, process), (power, knowledge), (power, process), (power, field), (process, location), (process, knowledge), (process, region), (process, process), (process, rule), (process, heuristic), (process, knowing), (act, location), (act, process), (act, rule), (action, location), (action, process), (basic cognitive process, higher cognitive process), (know-how, knowledge), (change, span), (power, content)	Score = 0.339 (6 pairs): (process, process), (method, Robotics), (change, span), (reduction, span), (reduction, Grasping), (method, Grasping)
Ruppin E (biochemistry and molecular biology)	Score = 0.599 (28 pairs): (psychological feature, psychological feature), (event, event), (event, phenomenon), (event, act), (knowledge, process), (power, process), (power, event), (process, process), (process, phenomenon), (process, rule), (process, act), (process, algorithm), (process, activity), (act, process), (act, event), (act, rule), (act, act), (act, activity), (action, process), (action, act), (action, activity), (method, system), (method, activity), (change, event), (change, phenomenon), (change, act), (reduction, preparation), (process, system)	Score = 0.332 (19 pairs): (event, event), (process, process), (act, act), (action, activity), (method, activity), (method, preparation), (method, planning), (method, programming), (change, preparation) (change, planning), (learning, planning), (change of magnitude, preparation), (reduction, act), (reduction, organization), (reduction, activity), (reduction, preparation), (reduction, planning),(reduction, programming), (action, planning)

Table 32.2 Performance assessment using OAEI 2012 benchmark

Test set	MLMA+			MLMA+ with depth weights		
	Precision	Recall	F-measure	Precision	Recall	F-measure
#101–104	0.74	1.0	0.85	0.93	0.84	0.88
#201–210	0.35	0.24	0.26	0.68	0.18	0.27
#221–247	0.71	0.99	0.82	0.94	0.66	0.75
#301–304	0.57	0.74	0.64	0.71	0.56	0.57
Average	0.59	0.74	0.64	0.82	0.56	0.62

constructs that make the concepts in the reference ontology generalized or restricted. Test set #221–247 contain test ontologies with structural change such as no specialization, flattened hierarchy, expanded hierarchy, no instance, no properties. The quality of both algorithms with respect to these two groups is quite similar since these modifications do not affect string-based and linguistic similarities which are the basis of both algorithms. Test set #201–210 contain test ontologies which relate to change of names in the reference ontology, such as by renaming with random strings, misspelling, synonyms, using certain naming convention, and translation into a foreign language. Both algorithms are more sensitive to this test set. Test set #301–304 contain test ontologies which are actual bibliographic ontologies.

MLMA+ with depth weights gives better precision but lower recall, thereby giving the average F-measure that is slightly less than that of MLMA+. MLMA+ discovers a large number of similar matched pairs whereas depth weights will decrease this number and hence precision is higher. But at the same time, recall is affected. This is because the reference alignments only lists pairs of terms that are expected to match. That is, for example, if the test ontology and the reference ontology contain the same term, the algorithm should be able to discover a match. But MLMA+ with depth weights considers the presence of the terms in the ontologies as well as their location in the ontologies. So an expected alignment in a reference alignment may be considered unmatched if they are near the root of the ontologies and are penalized by depth weights. The user-defined threshold *th* in the initialization phase of ontology matching can also affect precision and recall.

32.4.3 Discussion

Despite lower recall, we see that the concept of depth weights contributes something good to ontological profile matching since it can give a concise workable matching result. Both MLMA+ and MLMA+ with depth weights, however, do not consider the context of the terms so they can get a false-positive matching result if two ontologies contain a homograph. Another issue with the methodology is that research keywords are very technical and specific, and cannot be found in WordNet. We have to rely on the subject area or the generalized form of the keywords to form the ontology. Also, considering the noun phrase pattern or a tag count of a term is merely a way to resolve a problem although it may not give the most appropriate hypernym for a particular context. It is often the case that the specialized keywords are at the bottom of the ontology and all other terms built up by WordNet are more general and even abstract terms. It can be seen in Sect. 32.4.1 that the matching results are not very different when matching profiles of researchers in the same field, close fields, or different fields, in the sense that the research profiles mostly comprise general and abstract terms. Therefore the matching results contain matched pairs of such general terms. Our approach is in contrast to most related work which uses research-related information sources (e.g., personal and project Web sites) or relies on taxonomies of research areas. Although WordNet is not a database of research terminologies, we still see that it is a challenge to build research profiles from this rich source of information. It should be particularly useful for the research areas in which keywords are general terms and not very technical.

Since the methodology relies on lexical and semantic matching of terms in the profiles, any two researchers' profiles will match to a certain degree. Therefore, we recommend that the methodology be used when potential collaboration between two researchers is apparent.

32.5 Conclusion

This work explores the idea of an ontology-based methodology for building research profiles from ISI keywords and WordNet terms, and for finding similarity between the profiles. Relying on name similarity and linguistic similarity, the methodology can determine the degree of matching between the profiles as well as matched terms that represent similar research interests. We adopt the ODO algorithm for ontology building and MLMA+ algorithm for ontology matching and present a modification to MLMA+ with the concept of depth weights. A number of evaluations indicate that the methodology can give useful matching results.

For future work, further evaluation using a larger corpus will be presented. An experience report on practical use of the methodology on establishing research collaboration can also be expected. In addition, it is possible to adjust the ontology matching step so that the structure of the ontologies and the context of the terms are considered.

References

1. Okubo Y (1997) Bibliometric indicators and analysis of research systems: methods and examples. OECD Publishing, Paris
2. Kamsiang N, Senivongse T (2012) Identifying common research interest through matching of ontological research profiles, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct. USA, San Francisco, pp 380–385
3. Kamsiang N, Senivongse T (2012) An ontological analysis of common research interest for researchers. In: Proceedings of 8th international conference on computing and information technology (IC^{2}IT 2012), pp 163–168
4. Ontology alignment evaluation initiative 2012 campaign. Available: http://oaei.ontologymatching.org/2012/benchmarks/index.html
5. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) ArnetMiner: extraction and mining of academic social networks. In: Proceedings of 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD 2008), pp 990–998
6. Zhang J, Ackerman M, Adamic L (2007) Expertise network in online communities: structure and algorithms. In: Proceedings of 16th international world wide web conference (WWW 2007), pp 221–230
7. Punnarut R, Sriharee G (2010) A researcher expertise search system using ontology-based data mining. In: Proceedings of 7th Asia-Pacific conference on conceptual modelling (APCCM 2010), pp 71–78
8. Trigo L (2011) Studying researcher communities using text mining on online bibliographic databases. In: Proceedings of 15th Portuguese conference on artificial intelligence, pp 845–857
9. Yang Y, Yueng CA, Weal MJ, Davis HC (2009) The researcher social network: a social network based on metadata of scientific publications. In: Proceedings of web science conference 2009 (WebSci 2009)
10. ISI web of knowledge. Available: http://www.isiknowledge.com
11. An YJ, Geller J, Wu Y, Chun SA (2007) Automatic generation of ontology from the deep web. In: Proceedings of 18th international workshop on database and expert systems applications (DEXA'07), pp 470–474
12. WordNet. Available: http://wordnet.princeton.edu/
13. Alasoud A, Haarslev V, Shiri N (2008) An effective ontology matching technique. In: Proceedings of 17th international conference on foundations of intelligent systems, pp 585–590
14. Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33:31–88
15. Wordnet::Similarity. Available: http://sourceforge.net/projects/wn-similarity
16. Yang H, Liu S, Fu P, Qin H, Gu L (2009) A semantic distance measure for matching web services. In: Proceedings of international conference on computational intelligence and software engineering (CiSE), pp 1–3

Chapter 33
Automated Attention Analysis Across Brands and Cultures in Online Beer Marketing

Tomáš Kincl, Michal Novák, Pavel Štrach and Michal Charvát

Abstract This chapter presents an extended study focused on application of automated attention analysis in online marketing. The research question we are trying to address is whether automated tools can be used to depict differences between brand related websites of beer companies. Automated and quick comparison of websites from different markets and cultures might provide stimulating and instructive feedback and thus become an invaluable tool for online marketers. In spite of being exploratory in nature, the study and indicates that the automated tools instead of human-centered attention analysis could be an inexpensive yet relevant tool for brand site development.

Keywords Attention analysis · Automated tool · Cultural differences · Eye-tracking simulation · Online marketing · Web design

33.1 Introduction

Technology opens new perspectives or can substitute for human element. New technologies in modern marketing have stimulated emergence of several innovative interdisciplinary approaches. For example, measuring responses to marketing

T. Kincl (✉) · M. Novák · M. Charvát
University of Economics, Prague. Jarošovská 1117/II 377 01 Jindřichův Hradec,
Czech Republic
e-mail: kincl.tomas@gmail.com

M. Novák
e-mail: novak@michalnovak.eu

M. Charvát
e-mail: charvatmi@gmail.com

P. Štrach
Škoda Auto University, Tř. V. Klementa 869, 293 60 Mladá Boleslav, Czech Republic
e-mail: Pavel.Strach@skoda-auto.cz

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_33,

stimuli through technologies such as functional magnetic resonance imaging forms the base of neuromarketing [1] which is on the intersection of neuroscience and marketing. Automated eye-tracking represents an interdisciplinary connection between marketing and visual perception/cognitive psychology and is an integral part of visual marketing [2].

Traditional human-based eye-tracking has been a research method focused on monitoring the eye activity. Research in (then low-tech experimental) eye-tracking became a scientific discipline more than 100 years ago. Rayner [3] dates the early beginnings of the research back to the end of 19th century when many characteristics of eye movements were discovered. During and after the 1930s, more applied research has emerged, encompassing e.g. experimental psychology or behavioral theories. From 1970s more improvements, especially in eye movement recording systems, have been achieved. Additional accurate measurements included in a wide variety of eye-tracking systems are now easily available [4]. Eye-tracking methods are broadly used in many disciplines such as neuroscience, psychology, or computer science and have been utilized in several applied fields, including engineering and marketing/advertising [5].

Marketing researchers and practitioners use a variety of methods to evaluate consumer reactions (physical or physiological) to advertising stimuli [6]. Marketing research involves the eye-tracking methods to determine consumer's visual attention over advertisements—in order to follow and process a visual marketing stimulus, consumers move their eyes [7]. Eye movement consists of two different components: saccades are rapid eye movements focusing at a specific area, while the fixations are relatively still moments during detailed visual processing [8].

The sequence of saccades and fixations across the visual stimulus (such as advertisement) is called a scanpath [9]. Recording and analyzing a scanpath can reveal objects which "pop-out" from the image. The order of objects and the path between them outlines how the observer perceives each scene or image. Although costs of eye-tracking experiments are high, advertising clutter means that ad pretesting becomes crucial to ensure the effectiveness [10]. The use of eye-tracking is not limited to print only, evaluated advertisements can be tv spots or online objects as well [11].

Wide-spread use of eye-tracking in online marketing also brings in a call for cheaper and less time consuming alternatives. There are studies indicating high correlation between movements of user's eyes and mouse pointer [12]. The results are promising however, the accuracy has still yet to be elevated as there are differences between where users look and where they point the mouse when browsing the website [13]. Almost two-decades of research in neuroscience and natural vision processing resulted in automated systems which can simulate the human attention more accurately. Such systems produce similar results to a common eye-tracking study. Attention heat maps are comparable to eye-tracking maps and can be interpreted in a similar way. The comparison is shown in Fig. 33.1.

Captured snapshots of a website can be automatically analyzed on various features such as color, orientation, density, contrast, intensity, size, weight, intersection, closure, length, width and curve of displayed objects. Text, skin color

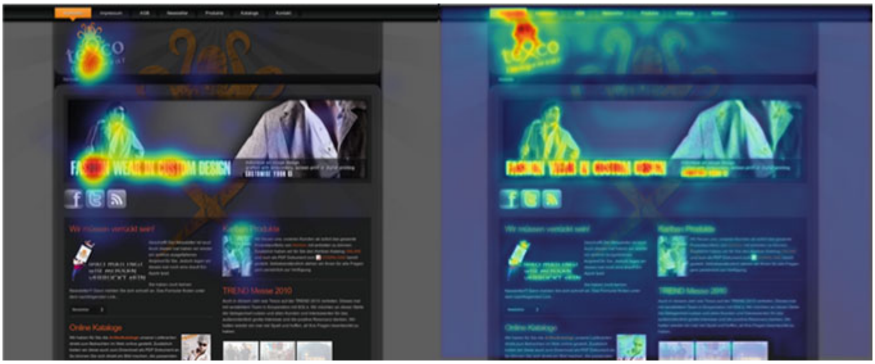


Fig. 33.1 Eye-tracking output (*left*) and automated attention analysis [63]

and face detection are also included [14]. Other tools may identify as many as 50 other features which attract human attention the most [15].

Automated tools for user-centered website assessment have been a recent phenomenon [16]. The most popular systems which allow uploading an image (e.g. screenshot) and get a visualization of estimated attention include eyequant (http://eyequant.com), Feng-Gui (http://www.feng-gui.com/) or Attention Wizzard (http://attentionwizard.com/). However, much debate [17] has evolved around their ability to simulate human perception and robust evidence is still missing. Authors of those automated tools claim 75–90 % correlation with real eye-tracking data [18]. The aim of this study is to contribute and expand [19] on this debate and to explore whether cultural differences in web design can be captured through automated eye-tracking tools. The study first discusses how the research was conducted, what results were achieved and final thoughts will be offered for further research inquiry.

33.2 Research Design and Methodology

The impact of culture on consumer behavior has been researched for decades. Market interdependence has prompted the emergence of theories, which attempt to explain differences between markets [20]. Studies of cultural differences introduced by Hall [21], Hofstede [22] or Trompenaars and Hampden-Turner [23] have become the classics of academic literature and the prominence of intercultural studies has survived or even grown in research interest [24] and there are new theories and approaches emerging (i.e. GLOBE study [25]).

Discussions about cultural specificity affected all components of marketing mix, including online marketing. Web designers began reflecting and adjusting to user's cultural characteristics [26]. Internet customers are different across the world according to their culture [27] however their online behavior can be similar in some ways [28]. Cultural differences between web designers and users determine success of web-based applications [29], however companies, origin of which is known, design their websites in a way that reflects their culture of origin rather than cultural norms of targeted customers [30]. Culture impacts on web design as well as on web content [31]. Websites need to be culturally adapted [32], although several cultures incline to adapt global patterns, behaviors or brands [33].

Literature highlights strong impact of high-and low-context communication styles on web design and website content organization and similar effects have been attributed to individualism-collectivism and power distance [32]. Websites in high context cultures display fewer interactive features [34], less varied and more shallow content [35] and are less informative [36]. High context websites are usually less clear and less attractive [32]. Websites from more individualistic cultures often offer more opportunities for face-to-face communication, because viewers feel more comfortable with impersonal communication [37]. Channels for mutual communication between clients and vendors are readily available in cultures with low power distance; the websites are rich on information as secretiveness would not be a source of advantage or power [34]. Evidence from studies focusing on the impact uncertainty avoidance on web design is ambiguous: the evidence is either inconclusive [29] or partially supportive for overall website attractiveness, content depth and information richness [32]. However, uncertainty avoidance influences the way users accept technological advancements [38].

Studies on cultural differences in web design confirmed that consumers from different cultures have different expectations and preferences about web design, which is also reflected by web design [39, 40]. This has been also confirmed by eye-tracking studies [41]. Therefore comes the research question:

> Could be the cultural differences in web design revealed by automated attention analysis tools?

Research sample consisted of commercial websites presenting beer brands. Alcohol consumption has been wide-spread across cultures, although motivation to drink alcohol differs [42]. Beer has been popular with consumers across countries and has been the most widespread alcohol drink [43]. Local beer markets offer opportunity to global as well as local brands, whereas brand cultural belongingness is usually well articulated. Beer is also a product with similar price and societal positioning in most cultures [44].

Beer consumption per capita has been the highest in Czech Republic [45] (however this small country ranks 21st in Beer Production Ranking [46]). Germany has been Nr. 2 in consumption per capita and is the 5th largest beer producer. Great Britain ranked 18th on the list of biggest per capita beer drinkers and is the 8th largest brewer in the world. These countries share strong beer culture and tradition [47], although national cultures differ [25]. Brazil represents Latin American culture [25] and also belongs to the group of the largest beer producers (comes 4th in beer production [46] and 27th in beer consumption per capita [45]). Japan represents a country with different approach to beer drinking [48]. Although the beer market has been well developed and beer drinking has become popular alongside of traditional liquors [49], Japan scores 35th in beer consumption ranking (but the 7th largest brewer absolutely). Culturally, Japan belongs to the Confucian group [25].

To establish a control group, in addition to selected markets, the study included a group of prime international brands with highest worldwide consumption (in case the brand featured on the local top consumption list as well, it was surveyed with the international group only as international brands could promote different than local cultural values). List of all websites included in the study is shown in Table 33.1. Selected websites were tested for similarity of user interface.

Opening introductory websites were surveyed unless there was a requirement to acknowledge user legal age first; in that case sites appearing after age check were analyzed. Opening website usually presents the key features kept for other subsites [50]. Analyzed websites were captured in 1,280 × 1,024 pixel resolution, which (or higher) was the most typical user resolution at the time of research [51]. Websites were then analyzed through Feng-Gui tool for the number of points of interest (AOIs) which would be likely the focal areas of user eye activity. The number of AOIs was captures as well as the overall area occupied by AOIs (in pixels). Analytical results can be visually displayed in a form of heat map (see example on Fig. 33.2).

Table 33.1 List of surveyed websites

Czech Republic [64]	Great Britain [65]	Japan [66, 67]
Gambrinus	Carling	Asahi
Radegast	John Smith's	Kirin
Staropramen	Old speckled hen	Suntory
Krušovice	Cobra	Sapporo
Pilsner Urquell	Newcastle brown	Orion
Budějovický Budvar	Fuller's London Pride	Baird
Bernard	Hobgoblin	Taisetsu Ji Bīru
Velkopopovický Kozel	Marston's Pedigree	Okhotsk Bīru
Starobrno	Abbot Alle	Tokachi Bīru
Ostravar	Tanglefoot	Otaru Bīru
Germany [68]	Brazil [69]	International [70]
Oettinger	Brahma	Snow (China)
Krombacher	Antárctica	Budlight (USA)
Bitburger	Cintra	Budweiser (USA)
Warsteiner	Bohemia	Skol (Brazil)
Beck's	Bavaria	Corona (Mexico)
Hasseröder	Nova Schin	Heineken (Netherlands)
Veltins	Kaiser	Miller Lite (USA)
Paulaner	Xingu	Guiness (UK)
Radeberger	Colonia	Coors (USA)
Erdinger	Stella Artois	Fosters (Australia)

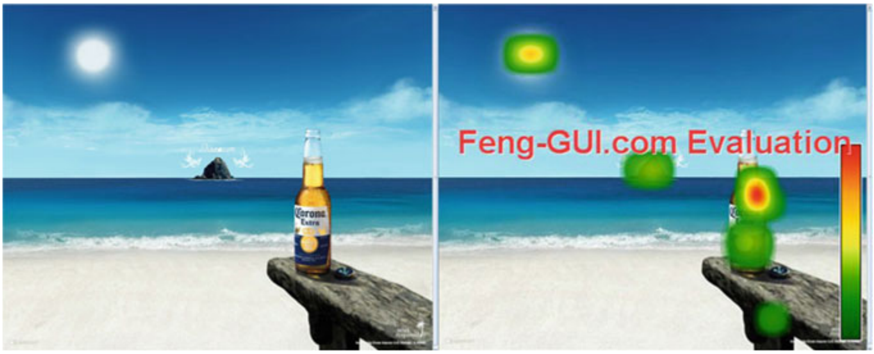


Fig. 33.2 Example of Feng-Gui analysis

It was expected that Japanese websites would be less visually appealing in comparison to Czech, German or British sites [32], less interactive [52] and offering less content [36]. International websites should than reflect and be more similar to presentations from countries with longer beer tradition. Most Czech brewers use a combination of dark colors, creating a feel of authority and seriousness, and one other color, which is either green (Radegast, Staropramen, Pilsner Urquell) sensing (in western cultures) nature and freshness, or brown/wine red (Gambrinus, Krusovice, Budweiser Budvar) intuiting health, earthiness or tradition [53]. Color choice has been consistent with long Czech brewing tradition. Websites use sizeable and illustrative pictures in vivid colors, in alignment with low context cultural habit [54]. Among top ten brands only one seems to take a different approach with combination of blue and white.

British and most German websites are similar to the Czech ones. Dominant colors refer to tradition, wholesomeness, earthiness, dependableness, steadiness, and healthiness [55]. Green or brown color is used quite often with the British sites (John Smith's, Old Speckled Hen, Newcastle Brown, Hobgoblin or Marston's Pedigree) and other markers of low context cultures are also present. Even though rich and vibrant colors are present on the German websites, the colors vary more from light hues (white, yellow—Erdinger, Bitburger, Veltins) to dark ones (dark grey and black—Warsteiner, Radeberger). The German websites resemble (in the variation of color usage) in that sense the group of Brazilian or International websites.

Brazilian websites use mostly warm and lighter colors. Among top ten brands only one takes a different approach and uses a dark color scheme (Xingu). Displayed motives are almost strictly limited to bottles of beer. There are no objects intuiting the naturalness, earthiness or healthiness on Brazilian websites which is another main distinction from Czech, British or German websites.

Color scheme of Japanese beer websites speaks a completely different language. The dominant color is white. In western countries, this would be the choice for sensing simplicity, neatness and precision. However in Asian countries white is perceived as color of death, mourning and spirit [53]. A little different are websites of Baird and Taisetsu Ji Beer, which remind websites of European brands and choose combinations featuring black or dark colors.

International websites are a diverse bunch of brands, originating from distant parts of the world. The first place goes to Chinese Snow, North America is represented by Budweiser, Central America by Corona; Heineken and Guinness hold European flags and Fosters comes from Australia. International brands push the feeling of uniqueness. Combinations of darker and another color are very frequent: Heineken—green, Budweiser—red, Fosters—blue, Guinness—blue. Other brands (Corona, Miller Lite, Coors Light, Bud Light) take a different route and prefer blue. Blue is a popular corporate color creating impression of trustworthiness. Very different is Snow website—its Chinese origin is reflected in color choice and dominant white reminds several Japanese brands.

33.3 Results and Discussion

Paired similarity in number and area of AOIs was tested through non-parametric Mann–Whitney test [56] which is deemed suitable for this type of observations. The analysis was performed through SPSS Statistical Software Websites from two different countries/groups were compared against each other. A complete picture depicting dyadic variations was compiled. Table 33.2 shows the key findings.

Difference in number of AOIs between the Czech Republic and Great Britain is not statistically significant (p value = 0.167) as well as the difference in AOI size (p value = 0.257). Czech and British beer sites are similar in terms of number and size of focal points. German websites are also similar in both surveyed aspects (number of AOIs p value = 0.303; AOIs size p value = 0.326) Czech, British and German beer websites seem to be a homogenous group in terms of these two focal attributes.

Although cultural classification [25] assign different cultural profiles to each nation, all three countries belong to the European Union and share the strong beer

Table 33.2 The results of AOI analysis

		Great Britain	Japan	Germany	Brazil	International
Czech Republic	AOIs	Same	Different	Same	Different	Different
	AOIs size	Same	Different	Same	Same	Same
	Great Britain	AOIs	Different	Same	Different	Different
		AOIs size	Different	Same	Same	Same
		Japan	AOIs	Same	Same	Same
			AOIs size	Same	Different	Different
			Germany	AOIs	Same	Same
				AOIs size	Same	Same
				Brazil	AOIs	Same
					AOIs size	Same

culture propped by lasting beer drinking tradition [47]. All three markets are labeled as low context [21]; hence approaching consumers through websites might be similar. All three website groups feature rich and vibrant colors with shades emphasizing trust, tradition or nature. Number of interactive objects is high complemented by rich information content. Such findings are in line with previous studies detailing low context online environments [52, 57].

Number of AOIs is significantly different for Czech sites and for the Japanese ones (p value = 0.032). The difference is significant even for the size of AOIs (p value = 0.007). The difference is not only in the number of AOIs (fewer of them in Japan) but in their size (larger ones in Japan). Japan is considered one of the most high context cultures. Absence of interactive features (such as consumer–consumer interactivity, fewer objects with smaller information content) could be explained through the lens of Japanese cultural distinctiveness [34, 35]. Whereas traditional beer cultures use rich and vibrant colors bringing forward serious, natural and calming impressions, Japanese beer websites utilize clean and plain white color which reduces first-impression attractiveness.

Number of AOIs is different for Brazilian and International brands and for Czech leading beers (p value = 0.004 for Brazilian group and p value = 0.008 for the International). However, the sites are not different in the area taken by AOIs (p value = 0.369 for Brazilian websites and p value = 0.406 for the International group). Czech websites were similar to International and Brazilian ones in the area occupied by AOIs but had higher amount of them. Several international brands do not originate from western low-context cultures but from Asia or Latin America featuring fewer interactive qualities [34, 52], less diverse and informative content [35] and are less informative in general [36]. Their main purpose may be intentional to cover large and diverse international audiences without targeting or offending any particular culture or adversely building on the global consumer.

There is no difference in number of focal elements (AOIs) and their size between British and German websites. P values do not reach the selected threshold (p value = 0.07 for number of AOIs; p value = 0.762 for AOIs size). As suggested before, Czech, British and German beer websites are a homogenous group in terms of selected attributes.

Number of AOIs is significantly different for British and Japanese beer websites (p value = 0.007). AOI size is different on the edge of statistical significance (p value = 0.059). The finding is consistent with the similarity found between Czech and British websites. Japanese websites contain fewer interactive options and fewer focal objects than the British ones.

Number of AOIs is significantly different for Great Britain—Brazilian and International brands comparison (p value = 0.001 for Brazilian and p value = 0.002 for International group). However, the screen area seized by AOIs does not seem to be dissimilar (p value = 0.364). British and International (Brazilian) brands are presented differently in terms of number of focal points but the overall space taken by key features is similar. Representation of non-Western brands from high context cultures (Asia, Latin America) which do not have English websites (or English language versions) could explain some of the differences.

The number of AOIs is higher on German websites and the AOIs are smaller than on Japanese websites. This corresponds with the results of the Czech and British group of websites. However, there is some variation in the number of AOIs and their size between Japanese and German websites, the analysis has not discovered any statistically significant difference between groups (*p* value = 0.266 for number of AOIs; *p* value = 0.131 for AOIs size). The situation is all the same when comparing German websites with Brazilian and International websites. Our study found also no statistical difference between the number of AOIs (*p* value = 0.380 for Brazilian group and *p* value = 0.175 for International websites) or their size (*p* value = 0.683 for Brazilian websites and *p* value = 0.705 for International group). German websites seems to be half way between traditional brewing countries (Czech and Great Britain) and the other groups.

The difference in number of AOIs is statistically insignificant for the analysis of Japanese and International (and Brazilian) brands (*p* value = 0.908 for International and *p* value = 0.835 for Brazilian group). In contrary, the difference is significant when comparing AOI size (*p* value = 0.023 for International and *p* value = 0.035 for Brazilian websites). Japanese and International (and Brazilian) websites are different in the screen area taken by AOIs (Japanese ones are bigger).

International (and Brazilian) websites stand on the edge between East and West. There was no statistical difference in number of AOIs (*p* value = 0.646) or their size (*p*-value = 0.806) between Brazilian and International websites. International and Brazilian brands apparently do not target consumers from traditional beer cultures specifically. It may well be that developing or emerging markets seem more promising for international beer brewers. Traditional beer cultures typically host several strong local brands which outperform international ones. International and also Brazilian websites present similar color schemes to traditional beer cultures, capitalizing on vibrant colors rather than on plain white.

33.4 Conclusion and Limitations

Cultural differences between websites can be deduced not only through demanding resource-consuming user testing or through expert panels. Cultural differences can be diagnosed via automated tools which simulate natural vision processing. Automated tools do not reflect local fluctuations or context and are prone to inappropriate sampling and personal bias. On the other hand, automated approaches offer less rich findings. The results could be also interpreted another way, since the automated tools do not perform testing on the same basis—user testing is based on specified task and eye-tracking results could be different according to different user scenarios. The results are also influenced by prior user experience or task with the website [58]. None of these factors are included in automated attention analysis and automated tools are not suited to depict such contingencies. Nevertheless, automated tools for website assessment have been a recent and increasingly popular phenomenon and have become prominent in other areas of web design. For instance, the CWS tool for harvesting visual cultural markers on the web [59] provides guidance for website layout and color scheme.

Automated tools can never fully substitute human experts in assessing human–computer interaction or user experience. Automated tools may provide fast and relatively inexpensive results for initial assessment of e-commerce and online marketing interfaces. The tools digest management-relevant outcomes for marketing managers or web designers who oversee and take decisions about the international portfolio of brand websites. Using initial automated testing may significantly reduce website development cost and contribute to more efficient marketing communications. The results of our study are consistent with studies suggesting that individualisms has been the key cultural dimension [32]. More individualistic cultures may need a greater number of eye catchers, impulses, whereas less individualistic cultures may focus longer on a smaller number of attention areas and rather explore interconnection between them.

Although the findings are stimulating, they should be interpreted with common scientific pre-caution. Correlation between traditional eye-tracking and automated eye-tracking has still been a matter of academic debate.

Our findings also may not reflect cultural differences only and could be based on designer's personal taste. Designer's personal bias is immanent to all website-based studies conducted outside closed labs and can be mitigated through inclusion of multiple websites residing in one cultural region. Cultural bias of automated tools also remains to be empirically verified. System calibration for other cultures would be useful feature and could enhance its predictive power. Further restrictions may rest on a relatively limited amount of output measurements—in comparison with rich findings from human eye-tracking [60, 61]. Automated tools usually offer just a handful of indicators: identification of areas of interests, their size and process of transition between various AOIs.

Future research could address some of the limitations; replicate our study on a set of different globally similar products, with real eye tracking or with utilization of multiple eye-tracking automated tools. Similarly, research in neuromarketing [1, 62] could become a viable avenue for future studies. Before any further work is done, academia as well as practitioners can utilize automated tools as an easy partial answer to a complex issue.

Acknowledgments This work has been supported by GACR grant no. P403/12/2175 and IGA VŠE 14/2012.

References

1. Lee N, Broderick AJ, Chamberlain L (2007) What is neuromarketing? A discussion and agenda for future research. Int J Psychophysiol 63(2):199–204
2. Wedel M Pieters R (2008) Eye tracking for visual marketing. Now Publishers Inc, USA
3. Rayner K (1998) Eye movements in reading and information processing: 20 years of research. Psychol Bull 124(3):372

4. Duchowski AT (2002) A breadth-first survey of eye-tracking applications. Behav Res Methods 34(4):455–470
5. Rayner K, Rotello CM, Stewart AJ, Keir J, Duffy SA (2001) Integrating text and pictorial information: Eye movements when looking at print advertisements. J Exp Psychol: Appl 7(3):219
6. Zikmund WG, Babin BJ (2006) Exploring marketing research. South-Western Pub
7. Wedel M, Pieters R (2007) A review of eye-tracking research in marketing. Rev Marketing Res 4:123–147
8. Buswell GT (1935) How people look at pictures: a study of the psychology and perception in art, University of Chicago press, Chicago
9. Noton D, Stark L (1971) Eye movements and visual perception. Scientific American
10. Berger S, Wagner U, Schwand C (2012) Assessing advertising effectiveness: the potential of goal-directed behavior. Psychol Marketing 29(6):411–421
11. Duchowski AT (2007) Eye tracking methodology: theory and practice. Springer, New York
12. Chen MC, Anderson JR, Sohn MH (2001) In: what can a mouse cursor tell us more? correlation of eye/mouse movements on web browsing, CHI'01 extended abstracts on human factors in computing systems, ACM, pp 281–282
13. Byrne MD, Anderson JR, Douglass S, Matessa M (1999) In: Eye tracking the visual search of click-down menus, ACM, pp 402–409
14. Mancas M (2007) Computational attention: modelisation and application to audio and image processing. Faculty Eng, Mons, Belgium
15. White matter labs understanding the world's most complex search engine: the human visual system. http://eyequant.com/science (accessed 06/27)
16. Kondratova I, Goldfarb I (2009) Cultural interface design advisor tool: research methodology and practical development efforts. Internationalization, design and global development. Springer Berlin Heidelberg, pp 259–265
17. Harty J (2011) Finding usability bugs with automated tests. Commun ACM 54(2):44–49
18. Feng-Gui How does it work? http://www.feng-gui.com/science.htm (accessed 06/27)
19. Kincl T, Novák M, Strach P (2012) In automated attention analysis: a valuable tool for online marketers?, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26, San Francisco, USA, pp 386–391
20. Maheswaran D, Shavitt S (2000) Issues and new directions in global consumer psychology. J Consum Psychol 9(2):59–66
21. Hall ET (1959) The silent language. Doubleday, New York, Vol 3
22. Hofstede GH (1984) Culture's consequences: International differences in work-related values. Sage Publications, California
23. Trompenaars F, Hampden-Turner C (1995) The seven cultures of capitalism: value systems for creating wealth in Britain, the United States, Germany, France, Japan, Sweden and the Netherlands. Judy Piatkus Publishers Ltd, London
24. Engelen A, Brettel M (2010) Assessing cross-cultural marketing theory and research. J Bus Res
25. House RJ (2004) Culture, leadership, and organizations: the GLOBE study of 62 societies. Sage Publications, California
26. Hsieh H, Holland R, Young MA (2009) Theoretical model for cross-cultural web design. In: Human centered design, Kurosu M (ed) Springer Berlin/Heidelberg, 5619, pp 712–721
27. Vuylsteke A, Wen Z, Baesens B, Poelmans J (2010) Consumers search for information on the internet: how and why China differs from Western Europe. J Interact Marketing 24(4):309–331
28. Brashear TG, Kashyap V, Musante MD, Donthu N (2009) A profile of the internet shopper: evidence from six countries. J Marketing Theor Pract 17(3):267–282
29. Baack DW, Singh N (2007) Culture and web communications. J Bus Res 60(3):181–188

30. Saffu K, Walker JH, Hinson R (2008) Strategic value and electronic commerce adoption among small and medium-sized enterprises in a transitional economy. J Bus Ind Marketing 23(6):395–404
31. Blake BF, Shamatta C, Neuendorf KA, Hamilton RL (2009) The cross-national comparison of website feature preferences: a practical approach. Int J Int Marketing Adv 5(3):145–165
32. Usunier JC, Roulin N, Ivens BS (2009) Cultural, national, and industry-level differences in B2B Web site design and content. Int J Electron Commerce 14(2):41–88
33. Kjeldgaard D, Askegaard S (2006) The glocalization of youth culture: the global youth segment as structures of common difference. J Consum Res 33(2):231–247
34. Cho CH, Cheon HJ (2005) Cross-cultural comparisons of interactivity on corporate web sites: the United States, the United Kingdom, Japan, and South Korea. J Adv 34(2):99–115
35. Würtz E (2005) Intercultural communication on web sites: a cross cultural analysis of web sites from high context cultures and low context cultures. J Comput Med Commun 11(1):274–299
36. Suh KW, Taylor CR, Lee DH (2007) Empirical classification of web site structure: a cross-national comparison
37. Liao H, Proctor R, Salvendy G (2008) Content preparation for cross-cultural e-commerce: a review and a model. Behav Inf Technol 27(1):43–61
38. Lee I, Choi B, Kim J, Hong SJ (2007) Culture-technology fit: effects of cultural characteristics on the post-adoption beliefs of mobile Internet users. Int J Electron Commerce 11(4):11–51
39. Callahan E (2005) Cultural similarities and differences in the design of university web sites. J Comput-Med Commun 11(1):239–273
40. Goyal N, Miner W, Nawathe N (2012) In cultural differences across governmental website design, ACM, pp 149–152
41. Cyr D, Head M, Larios H (2010) Colour appeal in website design within and across cultures: A multi-method evaluation. Int J Hum Comput Stud 68(1–2):1–21
42. Kuntsche E, Knibbe R, Gmel G, Engels R (2006) Who drinks and why? A review of socio-demographic, personality, and contextual issues behind the drinking motives in young people. Addict Behav 31(10):1844–1857
43. Ferreira MP, Willoughby D (2008) Alcohol consumption: the good, the bad, and the indifferent. Appl Physiol Nutr Metab 33(1):12–20
44. Dimofte C, Zeugner-Roth K, Johansson J (2010) In: local or global brand choice: do travelers really prefer global brands?, global brand management conference
45. Kirin institute of food and lifestyle report, Vol 33. http://www.kirinholdings.co.jp/english/news/2011/1221_01.html (accessed 1/8)
46. Kirin institute report (2009) global beer production by country in 2009, Vol 26. http://www.kirinholdings.co.jp/english/news/2010/0810_01.html#table2 (accessed 1/8)
47. Mäkelä P, Gmel G, Grittner U, Kuendig HÉ, Kuntsche S, Bloomfield K, Room R (2006) Drinking patterns and their gender differences in Europe. Alcohol Alcohol 41(suppl 1):i8
48. Wilson TM (2005) Drinking cultures: alcohol and identity. Berg Publishers, Oxford
49. Francks P (2009) Inconspicuous consumption: sake, beer, and the birth of the consumer in Japan. J Asian Studies 68(01):135–164
50. Leavitt MO, Shneiderman B (2006) Research-based web design and usability guidelines. In Washington, DC, Government printing office. Online at: http://www.usability.gov/pdfs/guidelines.html, Vol 1
51. W3Schools Browser display statistics 2010. http://www.w3schools.com/browsers/browsers_display.asp (accessed 8/1)
52. Cyr D (2008) Modeling web site design across cultures: relationships to trust, satisfaction, and e-loyalty. J Manage Inf Syst 24(4):47–72
53. Thorell LG, Smith WJ (1990) Using computer color effectively: an illustrated reference. Prentice-Hall. Upper Saddle River, USA
54. Simon SJ (2000) The impact of culture and gender on web sites: an empirical study. ACM SIGMIS Database 32(1):18–37

55. Hope A, Walch M (1990) The color compendium. Van Nostrand Reinhold, New York
56. Black K (2009) Business statistics: Contemporary decision making. Wiley
57. Singh N, Zhao H, Hu X (2005) Analyzing the cultural content of web sites: A cross-national comparison of China, India, Japan, and US. Int Marketing Rev 22(2):129–146
58. Pan B, Hembrooke HA, Gay GK, Granka LA, Feusner MK, Newman JK (2004) In: the determinants of web page viewing behavior: an eye-tracking study, pp 147–154
59. Kondratova I, Goldfarb I, Gervais R, Fournier L (2010) In: Culturally appropriate web interface design: web crawler study, pp 359–364
60. Poole A, Ball LJ (2005) Eye tracking in human-computer interaction and usability research: current status and future prospects. Encyclopedia of human computer interaction, 211–219
61. Jacob RJK, Karn KS (2003) Eye tracking in human-computer interaction and usability research: ready to deliver the promises. Mind 2(3):4
62. Reimann M, Schilke O, Weber B, Neuhaus C, Zaichkowsky J (2011) Functional magnetic resonance imaging in consumer research: a review and application. Psychol Marketing 28(6):608–637
63. White matter labs over 90 % of an Eye-Tracking study's accuracy—at less than 1 % of the time and cost. http://eyequant.com/accuracy#samples (accessed 06/27)
64. E15 Heineken pohltí Zlatopramen. http://www.e15.cz/byznys/heineken-pohlti-zlatopramen-18421/ (accessed 8/1)
65. The publican desert island brands. http://www.thepublican.com/story.asp?storycode=63086 (accessed 8/1)
66. Japan today Asahi retakes top slot in Japan's beer market. http://www.japantoday.com/category/business/view/asahi-retakes-top-slot-in-japans-beer-market (accessed 8/1)
67. Gloag D Japanese beer brands. http://www.united-nations-of-beer.com/japanese-beer-brands.html (accessed 8/1)
68. BrauWelt (2011) Inlandsabsatz der 56 größten Biermarken. BrauWelt, 151(14):418–420
69. Brazilian bubble overview: Brazil's beer market at a glance. http://brazilianbubble.com/overview-brazils-beer-market-at-a-glance/ (accessed 10/1/2011)
70. Jones D Snow jumps bud light to be world No 1. http://www.reuters.com/article/2009/03/31/us-world-idUSTRE52U60B20090331 (accessed 8/1)

Chapter 34
Inspiring Creative Minds

Anjum Zameer Bhat

Abstract Inspiring Creative Minds is an initiative and effort to reason a different perspective of teaching and learning to help students thrive and meet challenges and at the same time assist educationalists to impart education in an efficient and effective manner. In the predominant era of education, teaching and learning must transform to help students foster aptitudes to prosper in 21st century. Classroom tasks assigned to students at all levels of teaching and learning are extremely essential and it is mandatory for students to consistently follow and present assignments not only for recording commendable grades in exams but for holistic development of strong conjectural and pragmatic skills associated to a particular subject. Revising the curriculum topics is an important aspect of effective teaching and learning alongside pinpointing those topics which learners have not been able to comprehend because of numerous reasons. Effective teaching and learning necessitates improvements in lecture delivery which can bring enormous improvements to learners understanding of concepts. This research work proposes three innovative practices of teaching and learning that consider above aspects of teaching and learning and not only assist to deal with concerns related to teaching and learning but originate a competitive atmosphere which augments performance of apprentices and faculty members. Inspiring creative Minds is a teaching and learning approach which uses various software tools and IT infrastructure to provide ease and inspiration to students for better learning and participation in student centered learning process; in addition Inspiring Creative Minds is a guide to assist teachers for implementing and practicing student-centered teaching.

Keywords Artificial experience · BhatMeer Inspirational Model · Golden S for novices · Inspiring creative minds · Inspirational model · Innovative teaching practice

A. Z. Bhat (✉)
Department of Computing, Middle East College, Knowledge Oasis Muscat, 79Al Rusayl
124, Sultanate of Oman
e-mail: azameer@mec.edu.om

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_34,

34.1 Introduction

1. Golden S for Novices

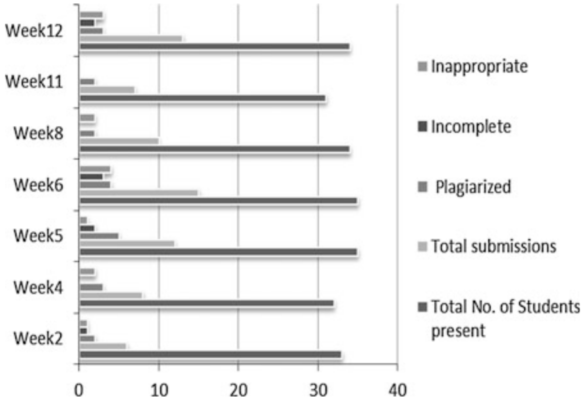
One of the well-known approaches in self-learning environment is to allow students to research at individual levels about a particular concept or a topic and show their work to the teacher for a realistic/honest feedback. The approach proves to be very beneficial for overall conceptual development of the students about a specific topic and can be equally helpful for the students' vis-à-vis module assessments. This type of student centered learning approach is used in various colleges and universities effectively and yields benefits to the students as they are able to research a concept on individual bases and can acquire knowledge beyond required level for a particular curriculum module.

However, due to the absence of any academic benefits like marks or grades, student participation is an area of concern. The above student-centered practice can yield benefits beyond expectation, if it is carried out properly with a motivating factor to encourage maximum participation of students in a particular cohort [1]. Students can research from multiple sources about a particular task or a concept and may achieve information more than what is illustrated in the module curriculum [1].

This helps students to comprehend concepts holistically which is essential for overall development of professional and academic competence. Figures 34.1 and 34.2 shows participation of students in the above student-centered learning environment in a cohort of 35 where students were assigned tasks at different times in an academic semester. In addition to less students participation other aspects which are counterproductive to the student's performance have also been taken into consideration.

Figures 34.1 and 34.2 clearly illustrates lack of interest and motivation in students to participate in a student-centered learning process apart from other issues of plagiarism, inappropriateness and incompleteness of tasks.

Fig. 34.1 Week-wise participation of students in student-centered learning



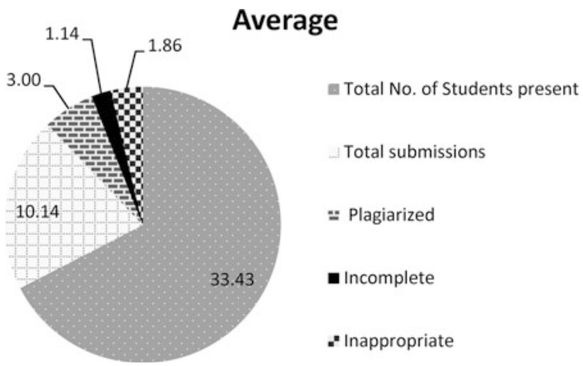


Fig. 34.2 Average participation of students in student-centered learning

When this activity was conducted on a group of 35 students out of which 33.43 students were on an average present in the class, only 10.14 students on an average submitted the assigned tasks, which of course is 28 % only and can be considered to be very less. In addition it is also evident from the figures, Figs. 34.1 and 34.2 that submitted assignments were found to be plagiarized, incomplete and inappropriate. Out of 10.14 submissions on an average 3 were plagiarized, 1.14 were incomplete and 1.86 submissions were inappropriate and overall more than 60 % of submissions were plagiarized, incomplete or inappropriate. In addition to very less students participation which was found to be 28 % only, if out of that also 60 % of the submissions are affected by various issues discussed above, it is useless to have this activity as it does not provide any benefits to the students and some mechanism should be deployed to either make certain modification to make this activity effective or it may be replaced by some similar activity proving extra beneficial for the students.

Human activities are driven by various types of motivations which can be autonomous or controlled motivations [2], One of the major reason of less student participation in student-centered learning environment is lack of appreciation and reward, students are not persuaded to work on any type of academic activity unless there are benefits or reasons in it, these *benefits/reasons* can be *academic (marks/grades)*, *appreciation, applaud, recognition, competition with the peers, honor, avoid indignity and so on.*

"Golden S for novices" is an innovative technique in which the performance of the students is closely monitored on regular basis and results of their assigned *tasks are displayed and showcased in a proper manner regularly in the classroom. Showcasing the results creates an intrinsic motivation* [2–4] *amongst the students* in a particular cohort to participate as they all can see performances of others. Those who are not performing and submitting their work, face a kind of humiliation as the results of their work are displayed to all in the classroom. On the other side, it *equally encourages* those students who are actively participating and performing well as they get *appreciation, applaud and recognition in the classroom*. The practice was tried on *6 cohorts of two different modules* and results were incredible as the average participation of students increased from 28 to 79.5 % [1].

Although the participation of students increased to an incredible degree but other issues like *plagiarism, inappropriateness and incompleteness of the assigned tasks worsened*. Maximum number of students submitted the *plagiarized, inappropriate and incomplete assignments*. A concept of "*Golden S*" was introduced to overcome this problem. Students were verbally informed that e.g. out *12 assigned tasks in an academic semester*, if they are submitting at least *10 tasks and out of 10 they get* "Good" *in five*, they acquire a "Golden S" and all these students will be given a *price/memento* at the end of the semester [5]. Students were coached well in understanding the preconditions for an assigned task to be "accepted" or marked as "Good". These preconditions are given as follows.

(a) *Submitted work should not be plagiarized.*
(b) *Incomplete work will not be accepted.*
(c) *Inappropriate work will not be considered.*
(d) *Students should use proper referencing in the submitted tasks.*
(e) *Student should be willing to represent/explain his work to the entire classroom if instructed by the teacher.*

"Golden S for novices" activity was carried out by using specialized custom designed software developed specifically for the activity [1]. The software developed provides an easy mechanism to input the performance of students during different weeks of an academic semester and provides a graphical display of individual progress of students. The software has also been designed so that less time is wasted while carrying out the activity of "Golden S for novices".

2. **Identify my Misdirected Teaches my Apprentice**

Revising the curriculum topics is an important aspect of effective teaching and learning alongside pinpointing those topics which learners have not been able to comprehend because of numerous reasons, henceforth revision of curriculum topics and *uncovering/identification* of *unproductive teaches* is utmost important for improved performance of students. "Identify my misdirected teaches my Apprentice" is an effective *student-centered learning activity* which provides a suitable mechanism to the students to revise their curriculum syllabus in a competitive learning environment alongside ensures identification of those topics which students have not been able to comprehend and empathize to a reasonable degree [1].

"Identify my misdirected teaches my Apprentice" is a simple but *effective innovative practice* in which a cohort is divided into equal groups of 2, 3, 4 or 5 members depending upon the strength of a particular cohort. The members of each group are chosen by the module instructor following the logic that each group should have a *combination* of *good, average and weak students*.

A question related to module curriculum is displayed on the screen for a specific member of a particular group. In case the respective member of the group answers the question *correctly*, *10 points* are given to that particular group. In case the respective member *fails to answer a question, -5 points* are given to the group with *Individual Presentation Penalty* (*IPP*) for the respective member. IPP means that the student who has failed to answer the question will have to prepare a presentation for the topic in question in the next class. Moreover same question is asked to the group to which the respective member belongs and in case none of the group members could answer the question, there are two choices provided for the group.

1. *Accept Group Presentation Penalty* (*GPP*) *with -15 points.*

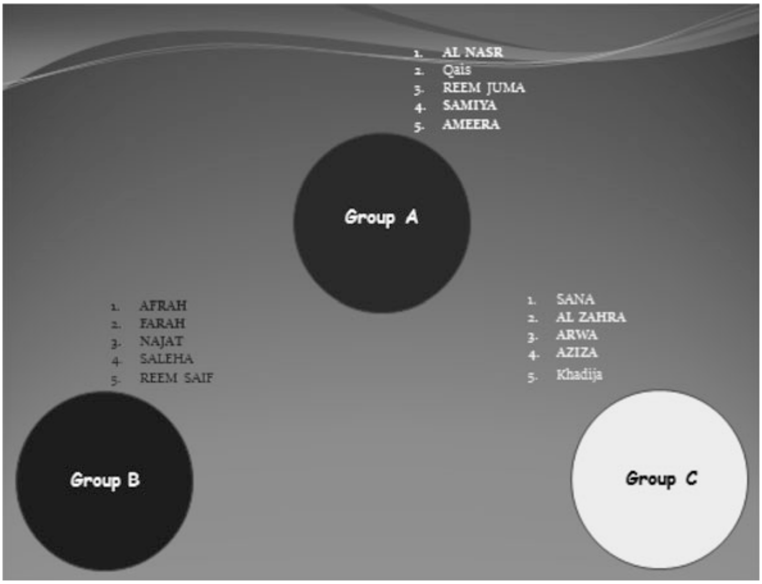
   Or

2. *Opt. for Repeat Class Penalty to the teacher* (*RCP*) *without any penalties to the group members or negative points to the group.*

Repeat class penalty can be accepted by the module instructor only in case *all the members* in *next group* could not *answer the question*, and if case be so Bonus for Identifying Misdirected Teaches (BIMT) *15 bonus points* are provided to the group for identifying misdirected teaches.

Example:—The below given example explains "Identify my misdirected teaches my Apprentice" activity. The example displays the activity conducted on a classroom of 15 students, divided into three groups. Group A, Group B and Group C.

Figures 34.3 and 34.4 show how the activity is carried out in the classroom. The group with highest number of points at the end is given *price/memento* as a token of appreciation; moreover the students who successfully answered all the questions asked to them also get a *price/memento*.



1. AL NASR
2. Qais
3. REEM JUMA
4. SAMIYA
5. AMEERA

Group A

1. AFRAH
2. FARAH
3. NAJAT
4. SALEHA
5. REEM SAIF

Group B

1. SANA
2. AL ZAHRA
3. ARWA
4. AZIZA
5. Khadija

Group C

Fig. 34.3 Identify my misdirected teaches my apprentice activity

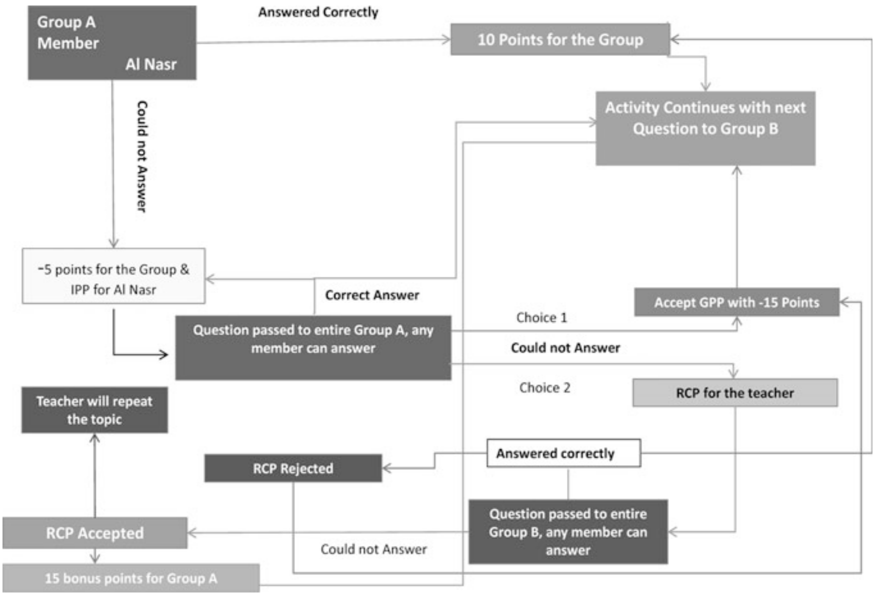

Group A
Member
Al Nasr
Answered Correctly
10 Points for the Group
Activity Continues with next Question to Group B
Could not Answer
-5 points for the Group & IPP for Al Nasr
Correct Answer
Question passed to entire Group A, any member can answer
Choice 1
Accept GPP with -15 Points
Could not Answer
Choice 2
RCP for the teacher
Teacher will repeat the topic
RCP Rejected
Answered correctly
Question passed to entire Group B, any member can answer
RCP Accepted
Could not Answer
15 bonus points for Group A

Fig. 34.4 Logical diagram of identify my misdirected teaches my apprentice

3. **Relate and Deliver with Artificial Experience**

Students tend to comprehend subject matter easily if it is thoughtfully presented to them [6]. Students always appreciate and commend teachers having enormous capabilities to correlate real life examples with the subject matter for better understanding and learning. Moreover teachers being the facilitators should be able to reduce the complexity involved in comprehending a particular concept. Although all educationalists at different levels of teaching and learning do provide students real life examples which most of the times are pictured and conceived by the teachers inside the classroom while delivering a lecture [1]; however precision in successfully correlating real life examples so that students can easily comprehend subject matter comes with experience and teaching a particular subject persistently for longer period of time, otherwise this approach can lead to chaos and confusion in student's mind with inexperienced faculty or a newly acquired subject matter [1]. Moreover even with experienced faculty members it cannot be expected that they can provide easy and suitable examples for every concept related to a particular subject. "Relate and Deliver with Artificial Experience" supports and recommends use of effective, easy and real life examples in a proper manner so that maximum outcome can be achieved.

Content-centered teaching focuses on and meets the requirements of the content [7, 8], *Instructor-centered teaching focuses on the teacher and teachers determine the content and organization of the courses as per their wishes and needs* [7, 8]. *Student-centered teaching focuses on student and content are largely determined*

Fig. 34.5 Process of authentication at an ATM

by student's needs [7, 8]. The below given examples shows the supporting content determined by student needs to explain the *Networking concept of Authentication, Authorization and Accounting.*

Example

Authentication

We are using Automated Teller Machines (ATM's) in our day-to-day life, when we go to a particular ATM for withdrawal of cash or balance inquiry we are actually logging on to the bank's network. ATM asks us a question "Who are you" and we reply it by inserting our *ATM card and entering pin (password)* as shown in Fig. 34.5, this process of identifying ourselves to the bank's network is called authentication. You need be a genuine user of the bank to log on to the banks network.

Authorization

After a successful authentication process you are logged on to the bank's network, where you can perform various operations like money withdrawal, balance inquiry, transfer of money and so on as shown in Fig. 34.6. Your authorization determines "What you can do", i.e. if you are having a balance of 100$ in your account you cannot withdraw 150$ from the ATM or you cannot transfer 200$ to some other account, moreover you can check the balance of your account not the balance of some other customer of the bank. Authorization determines your privileges and rights on a specific network.

Accounting

You might have also faced a problem that when you were trying to withdraw the money from the ATM, your account got debited but in reality you never received the money in your hand. You usually give a call to the bank's call center as illustrated in Fig. 34.7 and intimate them about the incident, but one important thing we need to notice is, where from the bank officials come to know whether

Fig. 34.6 Process of Authorization at an ATM

Fig. 34.7 Accounting

you received the money or there was some error which occurred during the transaction. Actually whatever actions you perform i.e. money withdrawal, balance inquiry, transfer or any other transaction, all of these actions are *monitored and archived* and different *logs/details* are available with the system for future reference. The Accounting refers to *storing of all the events which take place during a particular logon session which includes all successful and failure events*. Accounting determines "what you actually did" during a particular logon session on the network.

The above example if explained to the students as a supporting material before explanation of formal definitions for Authentication, Authorization and Accounting can help to a great extent to successfully convey the subject matter and students will feel substantial ease in comprehending the concepts. However there are certain holistic issues on which we need to have a minute analysis to make above practice much more beneficial for students as well as the teachers.

1. *Maybe there are better examples available for the above concepts which my contemporaries are using in their classrooms. Maybe there was a better example of the above concepts available with my predecessor who used to teach this module before me and now he is not a part of the institution.*
2. *A single teacher irrespective of his/her teaching experience and ability cannot conceive a good example for every concept he/she is teaching to his/her students.*
3. *Over a period of time when a teacher teaches a particular subject matter persistently, apart from getting a better hold of the subject, teacher also acquires many unique/useful ways and examples to convey the subject matter to his/her students in a better manner. Unfortunately all his/her experiences and examples which he/she had conceived while teaching a particular module extinct and are useless after he/she discontinues teaching that module or leaves the institution.*
4. *Newly recruited faculty members (with less experience) usually face problems in providing suitable and appropriate examples as they are still in developmental phase.*

Author provides a new concept of “Artificial Experience” which may prove to be very beneficial in addressing the above issues. “Artificial Experience” is a chronicled and documented experience and examples of a particular educationalist in a specific area of expertise. “Artificial Experience” may consist of documented examples which a particular teacher used over the years of his teaching career for some specific module or modules. “Artificial Experience” may also consist of recorded *audio/video aids* which will provide an idea of teacher’s elegance of teaching and delivery. These recorded and documented materials may be referred as “Experience Base”. The appropriateness and correctness of “Experience Base” can be assured by multiple reviews by experts. The author has conceived this idea from the well build concept of “Artificial Intelligence” [9].

34.2 Additional Resources Used

The innovative practices described above indeed require advanced resources to yield optimum results although first two practices i.e. “Golden S for Novices” and “Identify my misdirected teaches my apprentice” can be conducted using some basic application software, it is highly recommended to have a proper custom designed software to utilize the maximum efficiency, acquire desired outcome, avoid wastage of time and to automate the storage and retrieval of data and results. Author has developed software tools to support “Golden S for novices” and “Identify my misdirected teaches my apprentice”. These tools provide effective, easy and user-friendly interface to implement “Golden S for Novices” and “Identify my misdirected teaches my apprentice” in the classrooms. However these practices can well be implemented using basic applications like Microsoft word, Power point, excel etc. with some exertion which is proportional to the strength of a particular cohort. On the other hand we need to create a proper IT infrastructure to implement the concept of “Artificial Experience”. We need to have servers to host the “Experience Base” and accordingly specific services need to be enabled so that people can gain appropriate access to the resources as desired. Moreover it is reasonably apparent that “Experience Base” will not be developed within hours or days, it will certainly take some time and devoted efforts of various people in a particular institution. We can setup a team of persons consisting of one or more Departmental Heads, Senior Faculty members, members from IT support and administration who will be responsible for successful creation of “Experience Base” and implementation of “Artificial Experience”.

34.3 Related Work

Student-centered Learning is a methodology to education which emphasizes on the requirements of students rather than others involved in an educational process like teachers [10, 11]. Student-centered learning focuses on interests, aptitudes, requirements and learning approaches of students making teacher as a facilitator of learning [10, 11]. Student-centered learning allows students to actively participate in discovery learning processes. In the teacher-centered classroom, teachers are the primary source for knowledge, the focus of learning is to gain information as it is proctored to the student. On the other hand, student-centered learning is now the norm where active learning is strongly encouraged [10]. Students are now researching material pertinent to the success of their academia and knowledge production is seen as a standard [10]. The combination of a few learning practices such as *Bloom’s Taxonomy* [12] *and Howard Gardner’s Theory of Multiple intelligences* [13] can be valuable to a student-centered learning because they supports various modes of varied learning styles [12, 13]. Student-centered learning is essential as it encourages *discovery learning, promotes peer communication, enhances student motivation and enables students to take more responsibility for their learning* [10, 11].

According to *James Henderson* there are three basic principles of democratic living which are not yet established in our society in terms of education [6], these three basic principles which he calls three S’s of teaching are:

Self-Learning: Discovery learning processes from an autonomous viewpoint, engage oneself in generative process [6].

Social Learning: peer-to-peer interaction, collaborative thinking can lead to an abundance of knowledge [6].
Subject Learning: Subject matter thoughtfully prepared and presented [6].
Inspiring Creative Minds provides an effective mechanism to promote above principles of teaching and learning and inspire students to participate in student-centered learning process.

34.4 Results

1. **Golden S for Novices**

Feed-back on work or rewards lead to feelings of competence and so enhance intrinsic motivations [3], the technique proved to be extremely beneficial in reducing the issues of plagiarism, incompleteness and inappropriateness of work assigned to students. Apart from enhancing student's participation in a student-centered learning environment and reducing plagiarized, inappropriate and incomplete submissions, this technique also yields some hidden benefits. *Teacher is well aware about the performance and improvement of students at individual levels, teacher exactly knows which student has poorly performed in which assigned task, and teacher also knows which student has missed which particular topic in the classroom because of a skipped class*. The teacher can easily concentrate and allocate his attention to the students who are not performing to a satisfactory level and can arrange a separate meeting with them on one–one basis making "Office Hours" more meaningful for the teacher.

Figures 34.8 and 34.9 clearly indicate incredible increase in student's participation week 3 onwards and drastic decrease in plagiarized, inappropriate and incomplete work week 4 onwards.

Fig. 34.8 Week-wise participation of students after implementing Golden S for novices

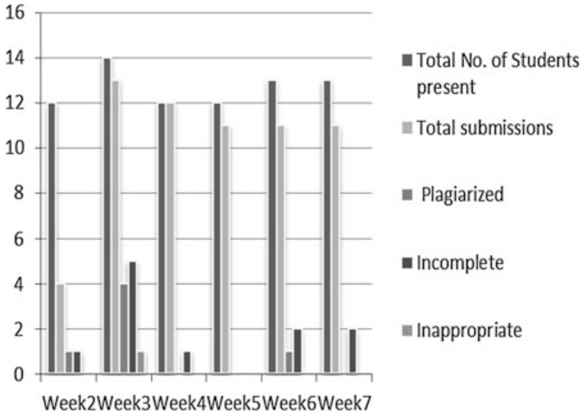
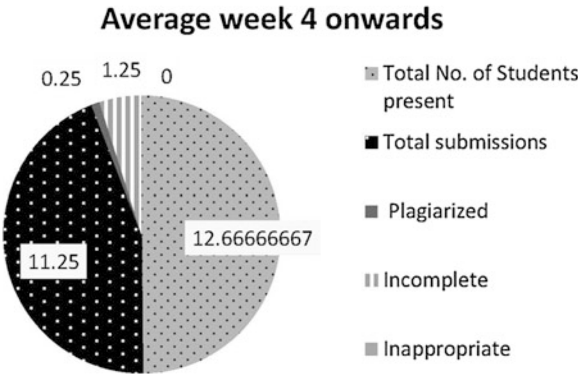


Fig. 34.9 Average participation of students after implementing Golden S for novices



2. **Identify my Misdirected Teaches my Apprentice**

Game2Learn [14], activity startlingly attained a huge response from five different cohorts on which it was practiced, a competitive environment was created in the classroom where students were trying their level best to perform better on individual levels as well as contribute to their respective groups. This activity increased interaction between the students and most of the groups and their respective members were preparing together to perform better in the activity. "Identify my misdirected teaches my apprentice" is an effective innovative student-centered learning practice which *promotes peer communication, enhances student motivation, enables students to take more responsibility for their learning and reduces disruptive behavior.*

3. **Relate and Deliver with Artificial Experience**

By using the concept of "Artificial Experience" we certainly can encourage and inspire teachers to take a more Student-centered approach of delivery for effective and successful Student-centered teaching and learning. Academic/classroom experiences of senior faculty members can be utilized in an efficient way by a less experienced faculty member to enhance lecture delivery, reduce complexity of topics so that students may be benefitted as they feel much ease in comprehending the curriculum topics with real life examples. The "Experience Base" proves to be very beneficial for new and less experienced faculty members as they acquire thoughtfully prepared supporting material and examples for the curriculum topics. "Artificial Experience" also consists of recorded *audio/video aids* which provide an idea of teacher's elegance of teaching and delivery, which provides new faculty members an idea of how they are supposed to deliver and conduct the classes. In addition the practice is very beneficial as the academic and classroom experiences of the faculty members who leave the institution or discontinue teaching a particular module are archived in a manner so that those can be reused for future purpose furthermore multiple faculty members handling different sessions of the same module can easily share their respective examples with each other and may adopt or use the examples of one another for the benefit of students.

34.5 Conclusion

Student-centered teaching and learning approach is being adopted worldwide in universities and colleges which has significantly transformed the education and helped students to foster skills in the modern era of education. Various innovative methods for student-centered teaching and learning have been proposed and practiced by the educationalists at different levels to benefit students however little work has been done to increase the participation and inspire students to actively involve themselves in student-centered teaching and learning especially in a learning environment where curriculum assessments are not student centered. Although we may not be able to modify curriculum assessments at various places due to restrictions from parent universities or affiliations, we still can certainly yield benefits of student-centered teaching and learning if we can encourage, inspire and motivate students to actively participate in student-centered learning and persuade faculty members to adopt student-centered teaching and learning approach. *Inspiring Creative Minds* is a mere effort to encourage the active participation of learners and at the same time convince and facilitate teachers to adopt student-centered approach for teaching and learning.

Acknowledgments I am thankful to Almighty Allah for bestowing me knowledge to accomplish this work. I thank HOD Computing Middle East College for his continued support and assistance. I am very thankful to all my contemporaries and friends who supported and encouraged me every time. I am greatly thankful to my family members especially my mother and my wife for their inspiration, support and guidance.

References

1. Bhat AZ (2012) BhatMeer inspirational model for student-centered teaching and learning, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Oct 2012, San Francisco, pp 297–302
2. Deci E, Ryan R (eds) (2002) Handbook of self-determination research. University of Rochester Press, Rochester
3. Deci EL (1975) Intrinsic motivation. Plenum, New York
4. Ryan RM, Deci EL (2000) Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. Am Psychol 55(1):68–78
5. Deci EL (1971) Effects of externally mediated rewards on intrinsic motivation. J Pers Soc Psychol 18:105–115
6. Henderson JG (1992) Reflective teaching: professional artistry through inquiry. Merrill Prentice Hall, Upper Saddle River
7. Professional Development Module: Student-Centered Teaching. By Vicky Lara, El Paso Community College. http://www.uottawa.ca/ http://www.uottawa.ca/ academic/ cut/options/ Dec_97/ Student_centered.htm
8. Sergio Piccinin. Making our teaching more student-centered at http://www.uottawa.ca / academic / cut /options / Dec_97 / Student_centered.htm

9. Jackson P (1998) Introduction to expert systems, (3 edn). Addison Wesley, Boston, ISBN 978-0-201-87686-4, p 2
10. Motschnig-Pitrik R, Holzinger A (2002) Student-centered teaching meets new media. Edu Technol Soc 5(4):160–172
11. Brush T, Saye J Implementation and evaluation of a student-centered learning. Educ Technol Res Dev 48(3):79–100
12. Bloom Benjamin (1956) Taxonomy of educational objectives. The classification of educational goals. Susan Fauer Company Incorporation, Chicago
13. Gardner H (1983) Frames of mind: the theory of multiple intelligences. Basic Books, New York
14. BARNES T, Powell E, Chaffin A, Lipford H (2008) Game2Learn: improving the motivation of CS1 Students. In: Proceedings of game development in computer science education (GDCSE'08)

Chapter 35
Mitigating Rural E-Learning Sustainability Challenges Using Cloud Computing Technology

S. A. Odunaike, O. O. Olugbara and S. O. Ojo

Abstract Internet technology is leading in transforming educational system by allowing different types of interactions among various educational institutions, all participating in global online innovations. In particular, educators have realized that technology enhanced learning, e-learning to be specific offers flexible and potent ways to accomplish wide spectrum of opportunities to relieve academic staff of excess workload and provide them with sufficient time to improve performance. Cloud computing technology can benefit educational institutions to provide human and material resources including course experts, digital contents, virtual laboratories and interactive classes, facilitate research, share knowledge, establish collaboration, support user mobility and perform computationally intensive laboratory experiments. However, extending cloud novelty and numerous benefits to rural e-learning raises many sustainability challenges. This work probes into how cloud computing can be effectively explored to mitigate rural e-learning sustainability challenges by utilizing descriptive research and scoping review approaches. The purpose is to raise awareness among stakeholders of educational institutions about prospects of using cloud computing. New issues of e-learning sustainability are discovered for future studies by considering focused areas of previous researchers and existing gaps. This work found energy and security as emerging sustainability issues in cloud computing applications to education.

S. A. Odunaike (✉)
Department of Software Engineering, Tshwane University of Technology,
Pretoria, South Africa
e-mail: odunaikesa@tut.ac.za

O. O. Olugbara
E-skills Institute Co-Lab, Durban University of Technology, Durban, South Africa
e-mail: oludayoo@dut.ac.za

S. O. Ojo
Faculty of Information and Communication Technology,
Tshwane University of Technology, Pretoria, South Africa
e-mail: ojoso@tut.ac.za

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_35,

Keywords Cloud · Education · E-learning · Internet · Rural · Sustainability

35.1 Introduction

The accelerated progress made in Information and Communication Technology (ICT) such as web technology, teleconferencing system and broadband internet has enabled educational institutions across the world to implement distance education to reach diverse population with easy access to resources without hindrances. This also has provided open learning environments for 24 hours day, 7 days a week and 52 weeks a year. Internet supported distance educational courses bring more than new students into online classrooms. In addition, they form a critical pressure point for challenging dominant assumptions and characteristics of existing traditionally organized twenty first century educational institutions.

There are proofs that e-learning provides educational system with innovative medium to effectively distribute and remotely share resources. In addition, e-learning provides excellent platforms for developing and organizing varieties of educational resources including class notes and outlines, long textual resources that resemble traditional textbooks, interactive non-linear tutorials, student questions and comments and even simulations of individual class sessions anytime and anywhere [1–4]. E-learning is enhancement and delivery of learning services using ICT and emerging educational technologies [5]. Research studies have equally shown existence of challenges and issues that threaten successful implementation and sustainability of e-learning, especially in rural settings [4, 6–9]. The same privileges and opportunities enjoy by educational institutions in urban settings to rake in e-learning for large economy of scale advantages cannot be said of rural educational institutions.

Cloud computing is a paradigm that focuses on sharing of resources and computations over a scalable interconnected nodes. Cloud computing can transform education and has following desirable properties that can be explored to solve inherent e-learning challenges. Dynamic scalability, self service, measured service, resource pooling, resource sharing, rapid elasticity, mobility support, service availability, fast connection, virtualization, multi-tenacity and pay as you consume [10–12]. In addition, cloud computing provides a great opportunity for faster processing power, cost effective maintenance, less computing downtime, large storage, maximum resource utilization, maximum return on investment, increase competitiveness, access to latest infrastructures and improves agility by allowing customers to provide products as utility services. Cloud computing is increasingly used for transacting business activities with lot of patronage from ICT based organizations. The continuous refinement of cloud computing by its providers increases possibility of making it an alternative technology for future investment. Cloud services are device, platform and location independent and can be accessed anytime and anywhere. Interactivity and accessibility with cloud technology is provided by principle of facilitating interaction between human and computer using application programming interface and software middleware that is used to aggregate geographically dispersed resources.

The promises of cloud computing are undoubtedly fascinating, but there are many sustainability challenges that almost render e-learning unusable, unsuitable and less effective in rural settings. Sustainability of e-learning initiatives is a common challenge across educational institutions regardless of project scale and focus [13, 14]. E-learning sustainability addresses current educational needs and accommodates continuous adaptation to change without receding in effectiveness [5]. It becomes apparent to investigate methods and technologies for sustainable e-learning as it is generally believed that e-learning cannot completely replace traditional classroom method. The required effort to adequately address e-learning sustainability has led us to the following research question. H*ow can cloud computing be effectively explored to improve sustainability of e-learning implementation in rural settings?* The outcome of sustainable e-learning studies will give understanding on how to adequately mitigate inherent challenges that have rendered it arduous to implement sustainable e-learning in rural settings. In addition, this can provide direction for selection and evaluation of future e-learning implementation strategies, create massive awareness among stakeholders and enhance competiveness. Section 35.2 presents discourse on literature covering cloud computing applications in educational domain. Section 35.3 discusses e-learning cloud to mitigate sustainability challenges. Section 35.4 presents scoping review result of the work. Section 35.5 concludes by suggesting recommendations for future rural e-learning cloud implementation.

35.2 Literature Review

ICT is a popular niche, but the understanding of important issues that determine influence of ICT on a particular organization remains elusive. The process that can allow smooth coordination of ICT and corporate strategy remains blurred. However, there are important opportunities arising from deploying ICT as a form of organizational internal strategy, business portfolio strategy and competitive strategy. Internal strategy is concerned with development of efficacious organizational structures and processes for achieving goals and objectives. Business portfolio strategy is concerned with choices of which industries to compete with and how to better position an organization in those industries. Competitive strategy focuses on competitive moves within industries in which an organization does business [15, 16]. Prospects of gaining competitive advantages have prompted quick implementation of e-learning as a formal mode of instruction delivery in higher educational institutions. This yields enormous amount of success to the extent of threatening traditional educational system to its value [2, 17, 18]. The same success cannot be claimed in rural educational institutions. Instead, it raises unpalatable sustainability challenges and issues that make e-learning almost unsuitable choice in rural settings [4, 6–9]. The quest for cost effective platforms to deploy ICT infrastructures and strategies to maximize profits and remain competitive drives rigorous search for paradigm shift in service deployment. This quest points to adaptation and advancement of cloud computing as a platform for utility service provisioning.

Cloud computing is an extension of traditional internet, service oriented architecture, web services and grid computing using virtual shared computing servers to deploy products, resources, software, infrastructures, devices, platforms and databases as utility services. The basic service ontology models of cloud computing are the following. Software as a Service (SaaS) provides opportunity for customers to run cloud applications through web browser, thin computing terminals and hosted desktops and eliminates the necessity to install and run these applications on consumer devices. Platform as a Service (PaaS) is the capability that allows customers to build and deploy specific applications using cloud software development environment (languages, libraries, functions, classes, components, services, packages and tools) supported by a cloud provider. Infrastructure as a Service (IaaS) provides customers with computational resources such as network (Network as a Service, NaaS) and data storage (Data as a Service, DaaS) to perform specific tasks. Computing as a Service (CaaS) is the capability that allows cloud providers to access raw computing power on virtual server. Hardware as a Service (HaaS) provides capability to access actual physical hardware and firmware as services. The deployment models of cloud computing include public cloud to provide services to customers through third party service provider such as Amazon. Examples of popular public cloud platforms are Amazon Elastic Compute cloud, Microsoft Windows Azure and Google App Engine. Private cloud is owned by an organization with a private network to benefit cloud computing behind firewall security. Community cloud is owned by several organizations for resource sharing. State cloud is owned by government with a government network to provide dedicated services to citizens. Hybrid cloud is a combination of private and public clouds for service provisioning.

A large number of studies on cloud computing focus on explanations of the technology, differences among similar technologies, security requirements, specific applications, future expectations, development, analysis and design [10, 12, 19–21] The applications of cloud computing in educational domain are emerging and have attracted attention of educators. Cloud computing can accelerate the adoption of different technological innovations in educational institutions. Researchers have provided a comprehensive introduction to applications and architectures of cloud computing in education [10, 14, 19, 20, 22–29]. In particular, cost analysis of cloud computing for education is presented [20]. An innovative ecosystem based on cloud and web 2.0 technologies are discussed [28]. E-learning cloud architecture is introduced to increase scalability, flexibility and availability of educational systems [26]. Education and Learning as a Service (ELaaS) is proposed for educational institutions with budget restrictions and sustainability challenges to use cloud formation best suited for their ICT activities

[14]. E-learning cloud architecture is introduced as a migration of cloud computing for future e-learning infrastructure [24].

35.3 E-learning Cloud for Sustainability

The promises of e-learning have recorded a slow growth in rural settings because of sustainability challenges, which can spell doom for making the technology as an instructional option. Cloud computing is an ideal technology for preserving numerous benefits of e-learning. In particular, e-learning cloud should be explored to mitigate sustainability challenges of connectivity, economical, energy, political, quality, resource, security, stakeholder, technological and training [30, 31].

The purpose of e-learning cloud is to provide E-learning as a Service (EaaS). This service model provides educational institutions with e-learning platforms, e-learning systems and e-learning resources as on demand services. The e-learning cloud model [22] proposes five components based on cloud computing ontology [32] to be dynamic datacenter, testing platform, security control, operational management and software platform. Dynamic datacenter is the physical heart of e-learning cloud to dynamically manage, deploy and secure services. Testing platform provides capability to incorporate new technologies and contents into e-learning platforms and allows educational institutions to test new e-learning applications prior to full launch. Security helps to control cases of server crashes, lost of data and applications located in remote sites. For instance, virtualization technology allows for rapid and cost effective replacement for a server. Operations and management function in e-learning platform is to provide services to third party such as educational institutions. Software platform that runs cloud, features several components including operating system, virtualization technology, resource management, e-learning platforms, collaboration, communication and security suites.

In addition, we propose that e-learning cloud provides efficient and dynamic peer-to-peer service discovery and roaming service component to fully harness benefits of cloud computing. Dynamic service selection and automatic discovery are important issues in cloud computing to simplify accessibility to services. At present, monitoring and discovering cloud services are unsatisfactory [33]. Cloud services should be accessible on a number of device interfaces, including web browsers, cell phone, smartphone, personal digital assistants, thin client such as CherryPal, iPad and tablet PC to enable ubiquitous accessibility to educational services and support for user mobility.

35.3.1 Connectivity Sustainability

Resilient and fast internet connection is a critical success factor for e-leaning cloud implementation. Internet connectivity is intermittently supplied and connection may be lost abruptly. The only way to participate in cloud computing is through fast internet connection and unlimited service access. Low connectivity rate as often experience in rural settings can spell a doom for e-learning cloud implementation. E-learning cloud service accessibility is inherently dependent on availability of internet connection. In addition, service quality heavily relies on connection speed that can require investment on network side [22]. Provisioning of complementary stable, high speed connectivity using efficacious communication technology such as broadband internet is required at affordable costs. The opinions of e-learning cloud providers may be sought on technology that can provide highly rated handshakes to equipments.

35.3.2 Economical Sustainability

Economic, capital, finance or cost is an important factor of e-learning sustainability. Embarking on a project of the magnitude of e-learning implementation in rural settings is capital intensive and can cause strains on budget. There is need to purchase the needed ICT equipment, network infrastructure, application software, computing server and allocate budget for system maintenance. Cloud computing is seen as a way to solve global economic crisis permeating educational institutions [21]. Staff and students can economically access diverse cloud resources through web pages and thin client interfaces to provide powerful functional capabilities and reduce costs of institutional expenses [19]. The cost of implementing cloud computing is low because customers need not purchase or own cloud equipments. Cloud service providers are paid to render resources to customers as on demand services. This arrangement has much to desire by allowing cloud customers to have full concentration on core businesses while cloud service providers bother with intrinsic cases, complexity, management of customers and infrastructures over cloud at reduced costs to customers. In most cases, risks and uncertainties associated with procurement of infrastructures, under-utilization of resources and low return on investment are shifted to cloud service providers. Consequently, e-leaning cloud shields customers from economic crisis because it provides access to services on demand. This mitigates ownership cost and money saved could be used to fund other sustainability needs.

35.3.3 Energy Sustainability

Most cloud service providers are located in urban settings where there is regular supply of energy. However, there is persistent, erratic and intermittent supply of energy in rural settings. The heighten service delivery protest is waged on erratic energy supply in urban settings. The energy provider even promise more black out in nearest future citing over loading, energy shedding and intense urbanization as excuses. This will remain an e-learning sustainability issue for as long as there is no alternative energy supply to rural settings. Alternative sources of energy will come at a high price for budget taking cognizance of steady maintenance [34, 35]. Energy efficiency is particularly considered critical for location based applications in mobile cloud computing [36]. Consumers of e-learning cloud services are mobile entities who carry mobile consumer devices such as cell phones, iPhone, iPod and iPads about for business transactions. Consequently, except consumer devices provide energy efficiency mechanisms to access e-learning cloud services, energy sustainability challenge will continuously remain a burden and research studies should adequately address energy efficiency challenge in cloud computing and mobile devices.

35.3.4 Political Sustainability

Diversities in culture, social, organizational policies and political affiliations are important role players in strategic decision making. An ugly political power play among stakeholders is detrimental to effective decision making and can widen e-learning sustainability gap. For instance, it is purely a political matter to use politics to advance issue of e-learning sustainability within educational institutions. A dynamic and constructive debates backed with due process of conflict resolution is important for any organization [37]. This will in turn filter out when e-learning cloud decisions are tabled for discourse among stakeholders. An organization where issues are meticulously debated based on their merits will be able to substantially influence support for e-learning cloud adoption and e-learning sustainability improvement decisions. The practice of e-learning at educational institution level cannot be successfully implemented and sustained without strong political leadership and full participation of all stakeholders.

35.3.5 Quality Sustainability

Educational service quality, educational attainment, performance, pedagogy and best practices are important issues of effective teaching and learning. The metrics of quality e-learning cloud implementation include student attainment, achievement and satisfaction of learning services, efficient service discovery, quality of service (fast connectivity, network reliability, service availability, currency and timeliness) and service level agreements. Rural e-learning offering must be competitive with those of urban settings by providing different kinds of interaction, enjoyment, excitement and satisfaction with new vista and pedagogy [17]. E-learning cloud can resolve quality issue, but it is advisable that choice of an e-learning cloud provider should be viewed with all seriousness. An e-learning cloud provider should be benchmarked with trusted providers and there must be constant review of service contract with expected industry requirements and standards. E-learning cloud providers should ensure that services are of high quality standards, reliable, scalable, uninterrupted, satisfactory and comparable with international standards of communities of practice and professional networks.

35.3.6 Resource Sustainability

Resource management is important for successful practice of education. E-learning service delivery must be supported with high quality of educational resources (experts, curricula and materials) as a prerequisite for effective e-learning implementation [38–40]. The e-learning model adopted should be complemented with availability of onsite qualified curriculum practitioners to assist teachers in developing and managing their course contents according to established e-learning standards. E-learning cloud allows for resource management and economy of scale, economy of scope, cost-effectiveness and efficiency practice by allowing for sharing and re-using of geographically distributed educational resources.

35.3.7 Security Sustainability

Privacy breach, security risks, distrust and cyber threats are still alive in cloud computing model [12, 41] because of the vulnerability nature of internet. Data stored in a public cloud can be exposed to malicious cyber crime attacks. In order to sustain e-learning cloud implementation, concerted efforts should be made not to compromise any form of cloud security and trust, including data protection against sabotage, destruction, natural disaster, frauds, worms and viruses. The attacks of malware and denial-of-service should be guided against by implementing sophisticated detection technology to proactively intercept any form of attacks. Although bulks of e-learning cloud infrastructures are on provider terrains, it is expected that adequate security should be ensured to guarantee e-learning sustainability. Due to seriousness of security, e-learning cloud security should not be down played where third party provider is the sole owner of infrastructure.

35.3.8 Stakeholder Sustainability

Stakeholder, that is administrator, teacher, researcher and student of educational institutions play important roles in decision making to facilitate effective implementation of e-learning in rural settings. Accommodating different stakeholder perspectives is a success factor for sustainable e-learning [13]. It is expected that stakeholders play important leadership roles and show strong commitment for e-learning sustainability issues. Although e-learning cloud may not have direct bearing on this factor, but decision taken to embark on e-learning cloud can serve as a proof of support to mitigate sustainability challenges. The intensity of impacts that e-learning cloud brings to educational reform can encourage stakeholders to show strong commitment to sustainability issues.

35.3.9 Technological Sustainability

The era of spending huge budget provision on procurement of state of art ICT applications and infrastructures are over because of emergence of cloud computing. Availability, affordability and usability of e-learning systems are important technological issues for sustainability. The lack of proper technology and curriculum required for e-learning adoption is a challenge facing many educational institutions. E-learning cloud models and cloud providers should ensure that customers have unlimited access to high level state of the art technologies to undertake core business activities of learning. This implies that technological sustainability challenge will be eliminated because of sophisticated technologies that are provided on e-learning cloud as services. In general, cloud computing comes with huge cost saving because technologies are provided as utility services and are not necessarily owned by individual customers.

35.3.10 Training Sustainability

E-learning cloud implementation offers a new pedagogical style for effective teaching and learning. Training of several teachers and students to acquire necessary knowledge, skills and competencies for professional development and innovation generally presents a challenge. A vital point of ensuring sustainable e-learning cloud is to invest in developing capacity of teachers to provide valuable knowledge, skills and competencies to integrate e-learning cloud into curriculum. Continuous training and professional development will create and foster communities of practice, professional networks and excellent engagement to guarantee e-learning cloud sustainability. An e-learning cloud provider, from time to time can be called upon to provide informative training and capacity development workshops for full utilization of the technology for effective teaching and learning.

35.4 Result and Discussion

In this section, we present result of scoping review methodological analysis of e-learning sustainability factors discussed by various researchers in research journals and conference proceedings published from the period of 2007–2012. In direction to gain a wider outlook, we used search engines to retrieve related research papers from databases of Springer LNCS (http://springer.com/lncs), ACM Digital Library (http://portal.acm.org), IEEE Explore (http://ieeexplore.ieee.org) and Google Scholar (http://scholar.google.co.in). The databases sufficiently cover the most related journals and conference proceedings within e-learning sustainability. Table 35.1 shows a set of search parameters and synonyms that were used to logically guide the search engines.

The scoping review methodological analysis scheme follows systematic review steps [42]. The analysis started with defining research area to be "the investigation of factors influencing e-learning sustainability". In the next step, we find status of the present work based on extant studies by identifying relevant publications from ACM, IEEE, Springer and Google databases. Irrelevant issues are eliminated in the next step according to papers that were judged obsolete and unconnected to the research area. The final step outlines research area by summarizing and interpreting findings. Contents of 22 papers found to be directly related to e-learning sustainability were analyzed. A score of 1 was allocated to a paper that discusses a particular e-learning sustainability factor of connectivity, economical, energy, political, quality, resource, security, stakeholder, technological and training. Table 35.2 shows the 22 research papers along with respective authors who have discussed a particular e-learning sustainability factor. According to the result in Table 35.2, the majority of research papers focused on economical factor (15.8 %) followed by resource (14.9 %). Quality factor contributed (12.9 %) papers, stakeholder (11.9 %), technological (11.9 %), political (10.9 %), training (8.9 %), connectivity (5.0 %), energy (4.0 %) and security (4.0 %).

The research papers where energy and security were discussed are related to e-learning cloud issues. These factors are classified into emerging and established categories. Emerging factors are those recently discussed by researchers as contributing to e-learning sustainability and they have low frequency of occurrence in papers reviewed. Established factors have high frequency of occurrence in many papers reviewed as contributing to e-learning sustainability. The classification procedure is based on unimodal iterative threshold selection algorithm, which is a one dimension variant of k-means algorithm [43]. The initial threshold of 10.0 % is taken as the mean of factors distribution and final threshold of 8.4 % was used for classification of factors. Emerging category has factors of connectivity, energy

Table 35.1 Searching parameters and synonyms

Parameters	Synonyms
Sustaining e-learning	Sustaining virtual learning; Sustaining online leaning
Sustainable e-learning	Sustainable virtual learning; Sustainable online learning
E-learning sustainability	Virtual learning sustainability; Online leaning sustainability
E-learning cloud sustainability	Virtual learning cloud sustainability; Online learning cloud sustainability

Table 35.2 E-learning sustainability factor, paper and year of paper publication

Paper	Years	1	2	3	4	5	6	7	8	9	10
Cook et al. [44]	2007								X		
Guglielmo [45]	2007		X		X	X	X			X	
Koohang and Harman [46]	2007		X			X	X		X		
Leary and Berge [47]	2007	X			X		X		X		X
Stephen [48]	2007		X			X	X				
Cartelli et al. [49]	2008		X			X				X	
Fuchs [50]	2008		X		X					X	
Robertson [51]	2008				X					X	X
Dong et al. [52]	2009		X			X	X				
Friesen [53]	2009	X	X		X				X		X
Stansfield et al. [54]	2009		X		X	X	X		X	X	X
Gunn [13]	2010					X	X		X		
Lai [55]	2010				X	X				X	
Stafford and Catlin-Groves [56]	2010		X				X				
Stepanyan et al. [5]	2010		X	X	X	X	X		X	X	X
Aljenaa et al. [22]	2011				X		X		X	X	X
Eswari [57]	2011		X				X	X	X	X	
Odunaike et al. [30]	2011	X	X	X	X	X	X	X	X	X	X
Toth [58]	2011	X	X			X	X		X	X	X
Koch [20]	2012		X	X		X	X	X			
Madan et al. [14]	2012		X								
Odunaike et al. [31]	2012	X	X	X	X	X	X	X	X	X	X

1 Connectivity, *2* economical, *3* energy, *4* political, *5* quality, *6* resource, *7* security, *8* stakeholder, *9* technological, *10* training

and security. Established category has factors of economical, political, quality, resource, stakeholder, technological and training.

Table 35.3 shows number of papers according to each factor of e-learning sustainability per year. The year 2011 signifies the period that e-learning sustainability challenges were mostly discussed (26.7 %). This is followed by 2007 (17.8 %), 2010 (15.8 %), 2012 (15.8 %), 2009 (14.9 %) and 2008 (8.9 %).

Figure 35.1 shows the graph of connectivity, energy and security to observe their trends from 2007 to 2012. It can be seen from the graph that interest in energy and security as e-learning sustainability factors is increasing from 2007 to 2012. This result implies that among the three emerging factors, energy and security are more likely to attract future discussion regarding e-learning cloud.

35.5 Conclusion and Future Work

In this chapter, we have evaluated the appropriateness of cloud computing for mitigating e-learning sustainability challenges. The technology is found worthy and ideal for mitigating e-learning sustainability challenges. The aim of raising

Table 35.3 Number of papers according to each factor per year

Factor/year	2007	2008	2009	2010	2011	2012
Connectivity	1	0	1	0	2	1
Economical	3	2	3	2	3	3
Energy	0	0	0	1	1	2
Political	2	2	2	2	2	1
Quality	3	1	2	3	2	2
Resource	4	0	2	3	4	2
Security	0	0	0	0	2	2
Stakeholder	3	0	2	2	4	1
Technological	1	3	1	2	4	1
Training	1	1	2	1	3	1

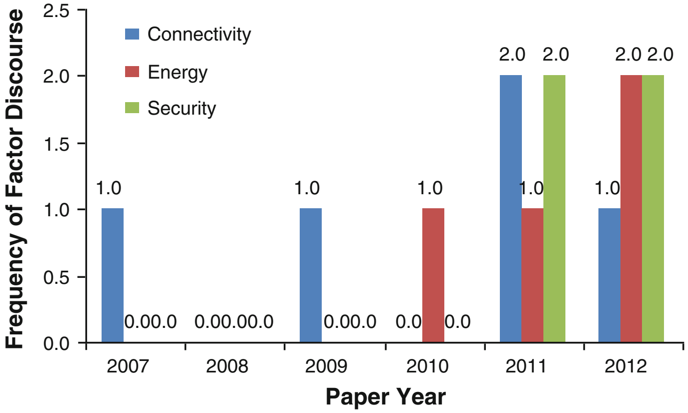


Fig. 35.1 Trends of discussion on connectivity, energy and security from 2007 to 2012

awareness on e-learning sustainability is to ensure that progress made by new reforms and activities of the nature, complexities and dynamics of e-learning are not derailed, but goes well into future. E-learning sustainability efforts explore possible ways of enduring the gains of e-learning implementation over time in the face of numerous challenges it presents to rural settings. These efforts have prompted us to look outwardly on cloud computing as a better way to implement rural e-learning.

Specifically, e-learning cloud offers the means of creating strong collaboration, reducing e-learning implementation costs, increasing competitive advantages, improving security at minimal costs, improving market flexibility, increasing return on investment, providing broad network access and high resource pooling. There are more benefits from e-learning cloud in terms of distributed provisioning of diverse kinds of educational services. However, we are expecting from cloud service providers, high quality rural educational services that provide energy efficiency and security measures using cloud computing and thin interfaces to access e-learning services. The business existence of e-learning cloud service providers will be at stake if ever there is breach of expectations. Conclusively, e-learning cloud should be used to mitigate sustainability challenges in rural settings. The future work will design and implement a robust energy efficient rural e-learning cloud application supporting efficient service discovery.

References

1. Byrne R (2002) Web-based learning versus traditional management development methods. Singap Manag Rev 24(2):59–68
2. Campbell N (2001) E-teaching, e-learning and e-education. A paper to inform the development of the ICT strategy in New Zealand for the Ministry of Education http://cms.steo.govt.nz/NR/rdonlyres/C11315DE-804A-4831-AB75-D4E77393DD8/0/eteaching.htm
3. Franklin S, Peat M (2001) Managing change: the use of mixed delivery modes to increase learning opportunities. Aust J Educ Technol 17(1):37–49
4. Holsapple CW, Lee-Post A (2006) Defining, assessing and promoting e-learning success: an information system perspective. Decis Sci J Innov Educ 4(1):67–85
5. Stepanyan K, Littlejohn A, Margaryan A (2010) Sustainable e-learning landscape: a scoping study (SelScope). The Higher Education Academy
6. Keegan D, Lossenko J, Mazar I, Fernandezs P, Paulsen MF, Rekkedal T, Toska JA, Zarka D (2007) E-learning initiatives that did not reach the targeted goals. MegaTrend in E-learning provision, NKI Publishing House, Bekkestua
7. McClelland B (2001) Digital learning and teaching: evaluation of developments for students in higher education. Eur J Eng Educ 26(2):107–115
8. Motiwallo L, Tello S (2000) Distance learning on the internet: an exploratory study. Internet High Educ 2(4):253–264
9. Wang LCC, Bagaka JG (2003) Understanding the dimensions of self exploration in web-based learning environment. J Res Technol Educ 34(3):364–373
10. Mathew S (2012) Implementation of cloud computing in education—a revolution. Int J Comput Theor Eng 4(3):473–475
11. Mell P, Grance T (2011) The NIST definition of clod computing. National Institute of Standards and Technology (NIST), Special Publication, pp 145–800
12. Singh S, Chana I (2013) Cloud based development issues: a methodological analysis. Int J Cloud Comput Serv Sci 2(1):73–84
13. Gunn C (2010) Sustainability factors for e-learning initiatives. ALT-J Res Learn Technol 18(2):89–103
14. Madan D, Pant A, Kumar S, Arora A (2012) E-learning based on cloud computing. Int J Adv Res Comput Sci Softw Eng 2(2) http://www.ijarcsse.com
15. Gerstein J, Reisman H (1982) Creating competitive advantage with computer technology. J Bus Strategy 3(1):53–60
16. Rockart JF, Scott Morton MS (1984) Implications of changes in information technology for corporate strategy. Interfaces 14(1):84–95
17. Bonk CJ (2009) The world is open: how web technology is revolutionizing education. Jossey-Bass, San Francisco
18. Christensen CM (2000) The innovator's dilemma: when new technologies cause great firms to fail. Harper Business, New York
19. Ercan T (2010) Effective use of cloud computing in educational institutions. Procedia Soc Behav Sci 2:938–942

20. Koch FL, Assuncao MD, Netto MAS (2012) A cost analysis of cloud computing for education. In: Proceedings of the 9th international conference on economics of grids, clouds, systems and services (GECON 2012)
21. Sultan N (2010) Cloud computing for education: a new dawn? Int J Inf Manage 30:109–116
22. Aljenaa E, Al-Anzi FS, Alshayeji M (2011) Towards an efficient e-learning system based on cloud computing. In: Second Kuwait conference on E-services and E-systems
23. Dodda RT, Smith C, van Moorsel A (2009) An architecture for cross-cloud system management. In: 2nd international conference on contemporary computing, vol. 40. Nioda, India, pp 556–567
24. Feng F (2010) Cloud-based IT infrastructure of next-generation telecom. Mobile Communications, No 8, pp 76–79
25. Ivanov II (2012) Cloud computing in education: the Intersection of challenges and opportunities. In: Web information and technologies lecture notes in business information processing, vol 101. pp 3–16
26. Masud AH, Huang X (2012) An e-leaning system architecture based on cloud computing. World Acad Sci Eng Technol 62:74–78
27. Mitchell P (2008) Learning architecture: issues in indexing Australian education in a Web 2.0 world. Indexer 26(4):163–169
28. Ouf S, Nasr M, Helmy Y (2011) An enhanced e-learning ecosystem based on an integration between cloud computing and Web2.0. In: Proceedings of IEEE international symposium on signal processing and information technology (ISSPIT), pp 48–55
29. Praveena K, Betsy T (2009) Application of cloud computing in Academia. IUP J Syst Manag 7(3):50–54
30. Odunaike SA, Chuene ND, Olugbara OO, Ojo SO (2011) Institutional e-learning sustainability for rural settings. In: Proceedings of the World congress on engineering and computer science 2011 (WCECS 2011), vol 1. San Francisco, pp 245–249
31. Odunaike SA, Olugbara OO, Ojo SO (2012) Using cloud computing to mitigate rural e-learning sustainability and challenges. In: Proceedings of the World congress on engineering and computer science 2012 (WCECS 2012), vol 1. San Francisco, pp 265–270
32. Youseff L, Butrico M, Silva DD (2008) Toward a unified ontology of cloud computing. In: Grid computing environments workshop
33. Goscinski A, Brock M (2010) Towards dynamic and attribute based publication, discovery and selection for cloud computing. Future Gener Comput Syst 26:947–970
34. Berl A, Gelenbe E, Girolamo MD, Giuliani G, Meer HD, Dang MQ, Pentikousis K (2010) Energy-efficient cloud computing. Comput J 53(7):1045–1051
35. Moghaddam FF, Cheriet M, Nguyen KK (2011) Low carbon virtual private clouds. In: 2011 IEEE 4th international conference on cloud computing, pp 259–266
36. Ma X, Cui Y, Stojmenovic I (2012) Energy efficiency on location based applications in mobile could computing: a survey. Procedia Comput Sci 10:577–584
37. Mallach EG (2000) Decision support and data warehouse systems. McGraw-Hill, New York, pp 57–65
38. Dabbagh N, Bannan-Ritland B (2005) Online learning: concepts, strategies and application. Pearson Merrill Prentice Hall, Upper Saddle River
39. Demirkan H, Goul M, Gros M (2010) A reference model for sustainable e-learning service systems: experiences with the join University/Teradata consortium. Decis Sci J Innovative Educ 8(1):151–189
40. Khan BH (2001) A framework for web-based learning. In: Khan BH (ed) Web-based training. Educational technology Publications, Englewood Cliffs
41. Zissis D, Lekkas D (2012) Addressing cloud computing security issues. Future Gener Comput Syst 28:583–592
42. Khan KS, Kunz R, Kleijnen J, Antes G (2003) Five steps to conducting a systematic review. J R Soc Med 96:118–121
43. Wikipedia. Thresholding (image processing) [Online] available: http://en.wikipedia.org/wiki/Threshoding_(image_processing)

44. Cook J, Holley D, Andrew D (2007) A stakeholder approach to implementing e-learning in a university. Br J Educ Technol 38(5):784–794
45. Guglielmo T (2007) A multidimensional approach to e-learning sustainability. Educ Technol 47(5):36–40
46. Koohang A, Harman K (2007) Advancing sustainability of open educational resources. Issues Informing Sci Inf Technol 4:535–544
47. Leary J, Berge Z (2007) Challenges and strategies for sustaining e-learning in small organizations. Online J Distance Learn Adm 10(3):1–8
48. Stephen D (2007) Models for sustainable open educational resources. Interdisc J Knowl Learn Objects 3:29–44
49. Cartelli A, Stansfield M, Connolly T, Magalhães H (2008) Towards the development of a new model for best practice and knowledge construction in virtual campuses. J Inf Technol Educ 7:121–134
50. Fuchs C (2008) The implications of new information and communication technologies for sustainability. Environ Dev Sustain 10:291–309
51. Robertson I (2008) Sustainable e-learning, activity theory and professional development. In: Proceedings ascilite Melbourne, pp 819–826
52. Dong B, Zheng Q, Yang J, Li H, Qiao M (2009) An e-learning ecosystem based on cloud computing infrastructure. In: Ninth IEEE international conference on advanced learning technologies, pp 125–129
53. Friesen N (2009) Open educational resources: new possibilities for change and sustainability. Int Rev Res Open Distance Learn 10(5):1–13
54. Stansfield M, Connolly T, Cartelli A, Jimoyiannis A, Magalhães H, Maillet K (2009) The identification of key Issues in the development of sustainable e-learning and virtual campus initiatives. Electron J e-Learn 7(2):155–164
55. Lai HF (2010) Determining the sustainability of virtual learning communities in e-learning platform. In: The 5th international conference on computer science and education. pp 1581–1586
56. Stafford R, Catlin-Groves CL (2010) Open source e-learning in higher education, problems, solutions and long-term sustainability of the approach. In: International conference of information society, pp 479–480
57. Eswari PRL (2011) A process framework for securing an e-learning ecosystem. In: 6th international conference on internet technology and secured transaction, pp 403–407
58. Toth KC (2011) An organizational approach for sustaining e-learning in a large urban university. In: Future of education conference.

Chapter 36
Information Harvest from Social Network Data (Facebook 100 million URLS)

P. Nancy and R. Geetha Ramani

Abstract Online social networks serves as an arena for its members to get in touch with each other, mutually share their information, ideas among themselves. In online social networks the members usually proclaim a profile, which consists of work and education, arts and entertainment and some basic information like gender, e-mail, etc., Such profile facilitates in spotting people, know about their interest, and interact with them in need. The intention of this research is to devise an algorithm to extract information such as name, email address, gender and interest of facebook users from a URL and to predict the gender if unspecified. The Dataset used in this work is a list of 100 million Facebook URLs. This research work paves a way to identify the email communities in Facebook. The outcome of this research reveals the fact that most of the email domains of the facebook user's fall into yahoo, hotmail, Gmail and msn. The other domains are with least number of users. The users with Yahoo id are higher when compared to other email domains. It also discloses that majority of the interest of facebook members is towards sports. It is followed by music, technology, travelling, God and Temple run, PC gaming.

Keywords Algorithm · Email address · Extraction · Facebook · Gender · Social network

P. Nancy (✉) · R. Geetha Ramani
Department of Information Science and Technology, College of Engineering, Anna University, Guindy, Chennai, India
e-mail: nancysundar09@gmail.com

R. Geetha Ramani
e-mail: rgeetha@yahoo.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_36,

36.1 Introduction

Online social networks are the one which pave way for various users to contact each other, give and take information and share their views among themselves. MySpace (over 275 million users), Facebook has more than 400 million users [1], Twitter has more than 40 million users are examples of wildly popular networks used to share among users. In online social networks the members usually announce a profile, which consists of work and education, arts and entertainment and some basic information like gender, e-mail, etc. Such profile information helps in identifying people, knowing about their interest, and interacting with them in time of need. However, in practice, not all users provide information about themselves. The profile of such people is said to be private. As per today's practice the members of the Facebook are asked to enter the profile information manually and it depends on the members, whether they wish to enter his/her details or avoids revealing the details [1]. The profile is said to be public if the information about the member is made public and it is private if the information is not revealed. In this chapter we propose a new Algorithm to retrieve the name, e-mail address and gender of a member from a URL. The Data set used in this research is 100 million Facebook URL which was hacked by Ron Bowes, an Internet Security Consultant [2].

In contrast, the goal of automatic information extraction (IE) is to discover relations between data items of interest and similar data items on a large scale and independently of their domain without any training [3, 4].

Facebook does not reveal a user's email address to any other user that is not in his friend list. In case the harvester is in the list, the user's email address is presented as a GIF image to prevent automated extraction. Twitter, on the other hand, does not reveal a user's email address in any form. However, the personal information that is revealed includes the user's name, personal web page, location and a short bio description [5].

The Web is an enormous source of information contained in billions of individual pages. Information extraction (IE) tries to process this information and make it available to structured queries. Most often, information extraction systems are targeted towards specific domains of interest and involve either manual or semi-automatic learning of the target examples involved [6]. The common format used by a web page is HTML. Data extraction from HTML is normally done with the help of wrappers. Existing Wrapper generation have the following features:

First, the wrapper generator works with information provided by the user or by external tool. Second, it is usually assumed that the wrapper works by knowing about the schema of data that is to be extracted. Finally, wrappers are generated by examining one HTML page at a time.

Another problem that prevails in the extraction of web page data includes the dominance of human factor (users) in the extraction process. In several similar applications such as RoboMaker (OpenKapow), YahooPipes, or Karma, that problem may occur because users search and select data table from a single web page manually. Since it is time consuming and costly, the process becomes less effective and efficient.

With respect to prediction of gender, the earlier approaches used the information provided by friends of a user based on user's affiliation in various groups. The accuracy of the prediction techniques was also low.

In this research we propose an algorithm for automatic extraction of data (name, email address and gender) from Facebook URLs and also combine the process of prediction of gender if unspecified in the user profile. The number of steps involved in the process of extraction of web information is less when compared to previous approaches. The proposed algorithm does not require Data Cleaning as the extraction process is highly accurate. The techniques used for prediction of gender include usage of first name of the user mentioned in the user profile.

36.1.1 Paper Organization

The chapter is organized in the following manner: Section. 36.2 gives a brief description of the related work. Section 36.3 narrates the proposed design of the work (overview of the system design, and the steps involved in the process), description of the Dataset. Section 36.4 explains about the experimental results obtained and projects the results obtained. Section 36.5 concludes the chapter.

36.2 Related Work

The work carried out so far by other researches that are related to retrieval of web information and prediction of gender is concisely presented here.

Gatterbauer [6] employed DOM (Document Object Model) as an approach for extracting web information and determining patterns from HTML tags or code structure in a web page. Gultom [7] used an approach to implement web table extraction and used the concept of mashing from HTML web pages by implementing the application they developed. It also used the concept of DOM generation for the HTML tags of the Web page.

Yanhong Zhai [8] proposed Partial Tree Alignment method which extracts data in two steps (1) Identifying individual data records in a page, and (2) Aligning and extracting data items from the identified data records.

Elena Zheleva [9] showed how an adversary can exploit an online social network with a mixture of public and private user profiles to predict the private attributes of users. Liu [10] used a Bayesian network approach to model the causal relations among people in social networks, and studied the impact of prior probability, influence strength, and society openness to inference the accuracy on a real online social network. Their experimental results revealed that personal attributes can be inferred with high accuracy especially when people are connected with strong relationships. Further, even in a society where most people hide their attributes, it is still possible to infer privacy information.

Hetherly [11] and his team explained how to launch inference attacks using released social networking data to predict undisclosed private information about individuals. They devised three possible sanitization techniques that could be used in various situations and explored the effectiveness of these techniques by implementing them on a dataset obtained from the Dallas/Fort Worth, Texas network.

Polakis [5] demonstrated how names extracted from social networks can be used to harvest email addresses. Cong Tang [12] and team developed a new and powerful technique for inferring gender for users who do not explicitly specify their gender. Having inferred the gender of most users in their Facebook dataset, gender characteristics were learnt and analysis on how males and females behave in Facebook was carried out. Different Gender prediction techniques like Offline Name List Predictor, Facebook Generated Name List Predictor, Local Information Predictor and Friend Information Predictor were designed and implemented individually. This research work has combined Offline Name List and Facebook Generated Name List for predicting the gender of a Facebook user [13].

36.3 Proposed Design of the System

This Section gives a brief description about the Dataset used for this research, the design and Architecture of the proposed System. The various steps involved in the algorithm for the process of extraction of information from the web page (Facebook user profile) are discussed in this section. The Techniques used in the Prediction of gender is also explained.

36.3.1 Dataset Description

The original dataset considered for this research is a torrent file downloaded from the blog of Skull Security. It was generated around July 15, 2010, by Ron Bowes, an internet security consultant [2]. He crawled the Facebook server of United States, and obtained the profiles of more than 100 million Facebook members. The dataset includes the Facebook URLs of various persons, the first name and last name of the users and the coresponding counts of the names as shown in Table 36.1.

Of the original torrent file, this research work focuses only Facebook-URLs. A Sample file of Facebook URL is shown in Table 36.2.

Table 36.1 Dataset description

File Name	Description
Facebook.rb	The script used to generate these files
Facebook.nse	The script that will be used for the second pass
Facebook-URLs	The full URLs to every profile
Facebook-names-original	All names, including duplicates
Facebook-names-unique	All names, no duplicates
Facebook-names-with count	All names, no duplicates but with count
Facebook-firstnames-withcount	All first names (with count)
Facebook-lastnames-withcount	All last names (with count)
Facebook-flast-withcount	All first initial last names(with count)
Facebook-first.l-withcount	All first name last initial(with count)

Table 36.2 A sample facebook URL

List of URL's
http://en-us.facebook.com/people/-/100000218612110
http://en-us.facebook.com/people/-/100000226945128
http://en-us.facebook.com/people/-/100000233424427
http://en-us.facebook.com/people/-/100000234406002
http://en-us.facebook.com/people/-/100000247916023
http://en-us.facebook.com/people/-/100000249924756
http://en-us.facebook.com/people/-/100000254263318
http://en-us.facebook.com/people/-/100000297669803
http://en-us.facebook.com/people/-/100000317949277
http://en-us.facebook.com/people/-/100000361441792
http://en-us.facebook.com/people/-/100000727219704
http://en-us.facebook.com/people/-/100000750793361
http://en-us.facebook.com/people/-/100000397174436

36.3.2 Overview of the System Design

The overview of the System Design is shown in Fig. 36.1.

36.3.3 Steps Involved in the Process

The algorithm designed is to implement the process of extracting the required data from the web page (Facebook User Profile). The required data which is to be extracted includes the user name, email address, gender.

The general flow of the work is briefly outlined as Algorithm and shown below.

Input: Text data *U*

Output: Database *D*

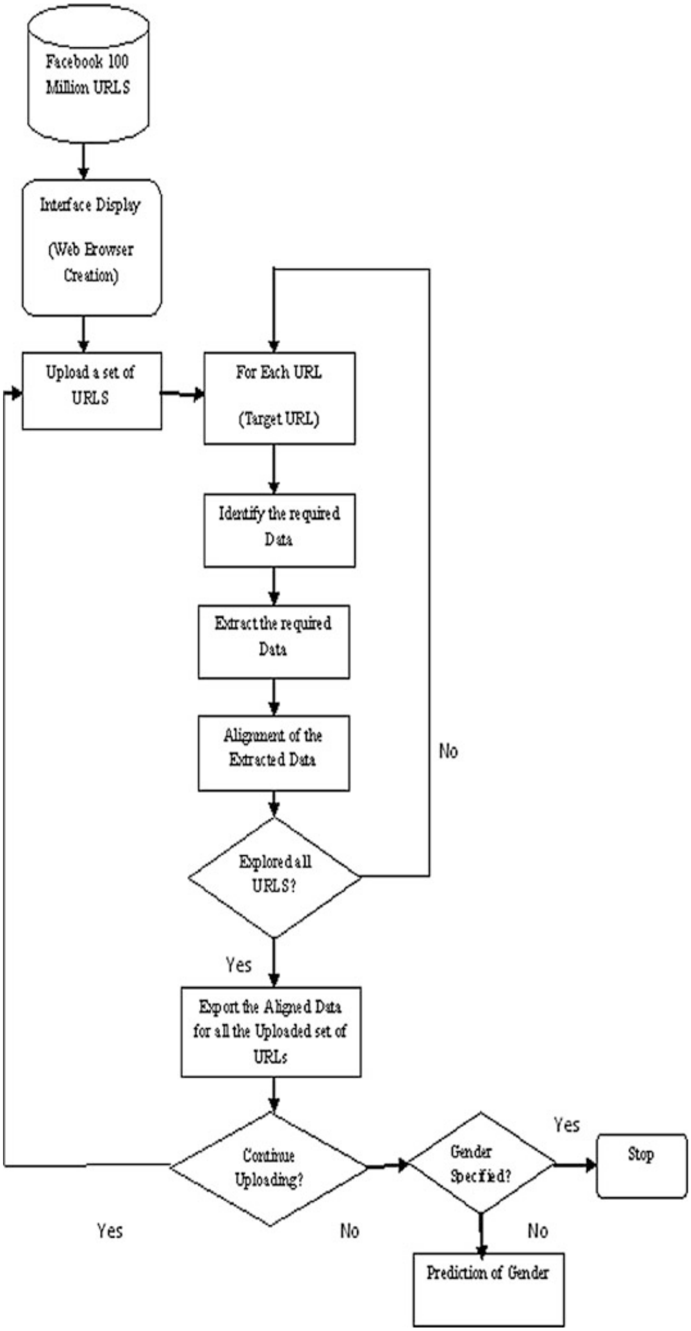

Facebook 100 Million URLS
Interface Display (Web Browser Creation)
Upload a set of URLS
For Each URL (Target URL)
Identify the required Data
Extract the required Data
Alignment of the Extracted Data
Explored all URLS?
No
Yes
Export the Aligned Data for all the Uploaded set of URLs
Continue Uploading?
Yes
No
Gender Specified?
Yes
Stop
No
Prediction of Gender

Fig. 36.1 System design

Parameters:

U Set of facebook URLS
D Structured database

Procedure

1. Read input data, U
2. Take a URL from the input data
3. Open URL through web browser
4. Extract html content of URL
5. Match keyword such as, name, gender, email, Interest, activities etc… To html document
6. Extract relevant data for every keyword
7. Put the extracted data to database, D
8. Go to step 2 until all the URLs processed
9. End.

In the above Algorithm ‘data’ denotes the required data which is to be extracted (Name, email address, gender, interests, activities etc.). Initially, a Web browser is created to get connected to Facebook.com. The Web browser also enables to implement the algorithm for extracting the data from web pages using graphical User interface. The web Browser is designed to enable creation of new session in Facebook.com, upload a set of URLs, to view the data extracted and to export the extracted data into an Excel Worksheet. In Step 1, a set of input data (URLs) is read. Step 2 deals with handling a single URL at a time. In step 3, the selected URL is opened through the Web Browser. Step 4 extracts the html content of the URL Step 5 proceeds with matching the keywords (required data) in the html. In Step 6, extraction of the required data is done and finally stored in a database. The Process is repeated until all the URL’s are processed. The original dataset contains about 100 million URLs which are voluminous to be loaded at once. Hence the URLs are divided into subsets to be loaded into the browser. From the set of URLs uploaded, the profile page of each URL is opened one by one to extract the required data.

36.3.3.1 Data Identification

For each URL loaded, the profile page is displayed in the web browser. In this step, html code of the loaded page is retrieved and stored. The code segment related to the retrieval of html code is as shown in Fig. 36.2.

Fig. 36.2 Code segment for retrieval of HTML page source

```
mshtml.IHTMLDocument2 doc =
(mshtml.IHTMLDocument2)_ActiveWebBrowser.Document;
Page code = ((mshtml.HTMLDocumentClass)(doc)).
                    documentElement.outerHTML;

str = Page code [13].
```

36.3.3.2 Data Alignment and Export

In the Data Alignment stage, the data extracted from each URL is viewed in the database. It can be used to view the extracted data immediately for the loaded page. In the Data Export stage, the listed sub items (extracted data) are exported into an excel sheet for further analysis and visualization. The Algorithm is shown in Fig. 36.3.

36.3.3.3 Gender Prediction

The extracted data exported into the excel worksheet in the previous step contains some unspecified values. The extracted data contains the name, URL, Email address and gender. Of all the extracted data, gender of a person can be predicted if unspecified and the prediction uses the algorithm shown in Fig. 36.4.

Popular baby names [14] is a first name list USA baby name list which consists of 1,736 male names and 2,023 female names.

Fig. 36.3 Algorithm for alignment and Export of data

```
Data alignment (URL, subitem)
Assign the Header of each subitem
Row=1;
While (Not EOF ()) For each URL uploaded
Add extracted subitem under its header
Increment Row
Open a New Excel Application
Open a New Worksheet
Row = 1;
Column =1;
For each (header in data alignment)
Column=1;
For each subitem
Worksheet cell (Row, Colum) = subitem;
Increment column;
Increment Row
```

Fig. 36.4 Algorithm for gender prediction

```
Gender Prediction ($name, $gender)
For every unspecified gender for $name
Repeat
If $name appears in the worksheet
Assign Gender with high probability
Else if $name €popular baby names
Assign Gender with high probability
Else if $name € Facebook namelist
Assign Gender with high probability
Until the all the unspecified gender is predicted [13].
```

Facebook namelist [15] also list the first name with a count of male and female Throughout the Gender Prediction Algorithm, the Probability was calculated using Bayes Theorem. The idea is

$$\begin{aligned} \mathrm{P(A)} &= \text{proportion of trials producing outcome as Male} \\ \mathrm{P(B)} &= \text{proportion of trials producing outcome as Female} \end{aligned}$$

If we consider only trials in which A occurs, the proportion in which B also occurs is P(B|A). If we consider only trials in which B occurs, the proportion in which A also occurs is P(A|B). In simpler form,

For events A and B, provided that $\mathrm{P(B)} \neq 0$

$$\frac{\mathrm{P(A|B) = P(B|A)P(A)}}{\mathrm{P(B)}}$$

Similarly P(B|A) is found. The higher of the two is selected for predicting the Gender [13].

36.4 Experimental Results

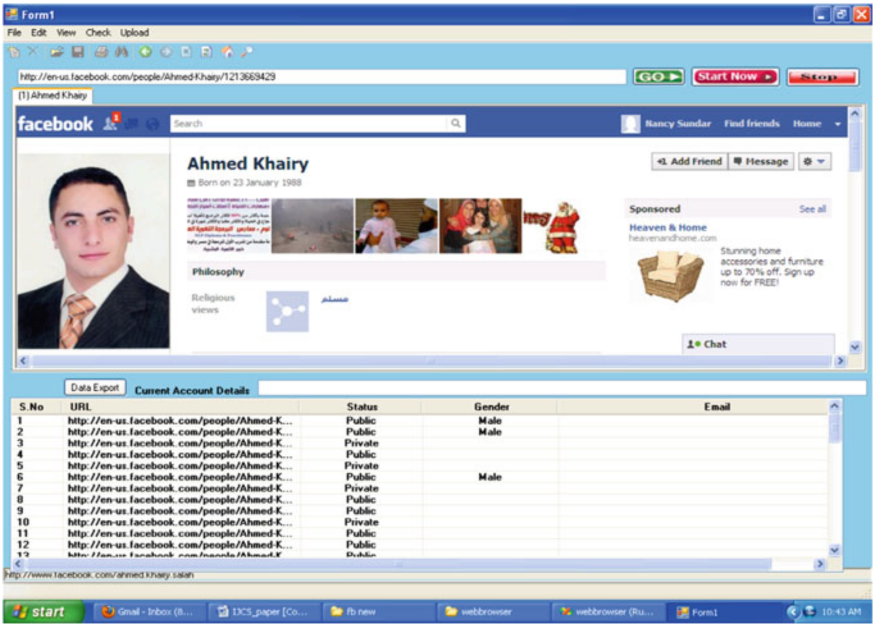
The entire application was developed using java as it has many inbuilt features useful for extraction of the data from the web. The Original Dataset contained 100 million Facebook URLs which is about 1.65 GB.

Text file Splitter is used to Split the original Data (100 million Facebook URLs) into Small Text files. The size of the Splitted file was about 1.35 MB with about 25,000 URLs in each file. The total number of Splitted files was about 1,232 text files which contained only 30 million URLs. This part of the research, considers only 30 million URLs for information extraction and the results obtained are pertained to 30 million URLs only.

The snapshot of the application with a web Page loaded for a specific URL with some of the extracted data is shown in Fig. 36.5.

The extracted data is properly aligned under specific headers. The data has some unspecified values for gender which are to be predicted. All the extracted data is exported into a excel worksheet for further analysis. Out of 30.82 million URLs, 7.93 million URL's were found to be under private category and 22.8 million URL's were found be under public category. From 22.8 URL's, Gender information was present in 19.45 million URL's. However, 13,311 URLs' contained email address. From the extracted email addresses, it was found that users belong to various domains like yahoo, hotmail, Gmail and msn. 6,004 users were found to be under yahoo domain, 2,792 in hotmail, 2,143 in Gmail and 2,372 in msn [13].

The majority of email domain communities which persisted in Facebook ULRs are shown in Fig. 36.6.



http://en-us.facebook.com/people/Ahmed-Khairy/1213669429

Current Account Details

S.No	URL	Status	Gender	Email
1	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
2	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
3	http://en-us.facebook.com/people/Ahmed-K...	Private		
4	http://en-us.facebook.com/people/Ahmed-K...	Public		
5	http://en-us.facebook.com/people/Ahmed-K...	Private		
6	http://en-us.facebook.com/people/Ahmed-K...	Public	Male	
7	http://en-us.facebook.com/people/Ahmed-K...	Private		
8	http://en-us.facebook.com/people/Ahmed-K...	Public		
9	http://en-us.facebook.com/people/Ahmed-K...	Public		
10	http://en-us.facebook.com/people/Ahmed-K...	Private		
11	http://en-us.facebook.com/people/Ahmed-K...	Public		
12	http://en-us.facebook.com/people/Ahmed-K...	Public		
13	http://en-us.facebook.com/people/Ahmed-K	Public		

http://www.facebook.com/ahmed.khairy.salah

Fig. 36.5 Snapshot of a loaded URL

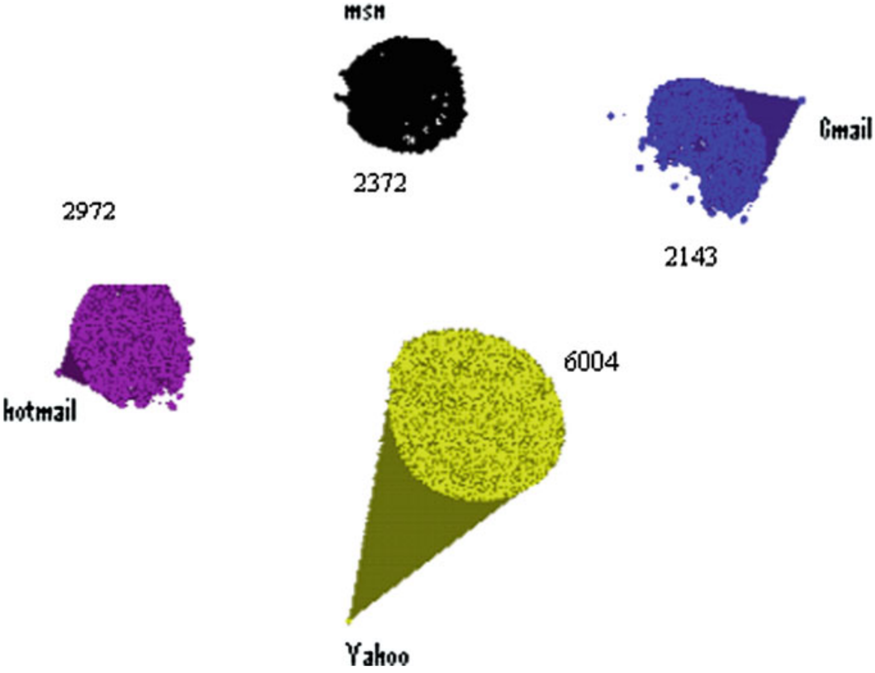


Fig. 36.6 Major email domain communities in facebook

Table 36.3 Top 10 interested activities identified from facebook users

S.No	Interest
1	Sports
2	Music
3	Technology
4	Travelling
5	God and temple run
6	PC gaming
7	Sociology
8	Driving
9	Dance
10	Sleeping

Out of 19.45 million URLs in which Gender Information existed, it was found that 11.47 million users were male and 7.97 users were female. In 3.45 million URLs, the Gender information was not specified. We used a combination of Gender Prediction techniques (Name centric approach) like popular baby name list, facebook name list to predict the Gender of 3.45 million URLs. From the predicted Genders, 1.85 million were predicted to be male and 1.57 million were predicted to be female [13].

From 22.8 million Public URL's only 5.36 million URL's contained favorites specified. In the favorites section many of the Facebook users have shown their interests in various fields, of which Sports gained the top most ranking. Nearly 18 lakh Facebook users are interested in Sports. Out of various sports, the highlighted games identified were Football, Basketball, Snooker, Squash, Pingpong, Tennis and cricket. Out of 18 lakh facebook users interested in sports, nearly 60 % were male. Around 6 lakh users were interested in Technology, 8 lakh users were fascinated in Music, 5 lakh users were involved in Travelling. Some of the users were concerned in God and Temple run, few were engrossed in Sociology and Driving.

The top 10 interests identified from facebook users are shown in Table 36.3.

36.5 Conclusion

In this research work, a new algorithm for extracting information from URL's (web) was proposed and executed. The Dataset used in this research is Facebook 100 million URL's. The dataset is very huge and hence we concentrated on only 30 million URL's as initial step. Any data required which is available in the URL can be extracted without any human intervention and stored in a database. This work focused on extracting name, gender, email and favorites (interests) and also predicts the gender if unspecified in the URL. It is evident from the data obtained, that the major email community of facebook users includes yahoo, hotmail, gmail and msn. It can be concluded that only 0.25 % of users have specified email address in their profiles. The top 10 interests among the facebook users were identified, of which Sports ranked the top followed by music, technology, travelling, God and Temple run, PC gaming etc… The chapter also uses various offline gender prediction techniques and predicted the gender of user. Thus it is apparent that any information can be extracted automatically from URL's.

References

1. Facebook statistics [Online] Available: http://www.facebook.com/press/info.php?statistics
2. Facebook 100 million user profile [Online] Available: http://www.skullsecurity.org
3. Mislove A, Viswanath B, Gummadi KP, Druschel P (2010) You are who you know: inferring user profiles in online social networks. In: Proceedings of WSDM, 2010
4. Chun H, Kwak H, Eom Y-H, Ahn Y–Y, Moon S, Jeong H (2008) Online social networks: sheer volume vs social interaction. In: Proceedings of IMC, 2008
5. Polakis I, Kontaxis G, Markatos E (2010) Using social networks to harvest e-mail addresses. In: Proceedings of WPES'2010
6. Gatterbauer W, Bohunsky P, Herzog M, Krupl B, Pollak B (2007) Towards domain independent information extraction from web tables. In: Proceeding of the international world wide web conference committee (IW3C2), May 8–12 2007, ACM, Banff, Alberta, Canada, pp 71–80
7. RAG Gultom, RF Sari, B Budiardjo (2011) Proposing the new algorithm and technique development for integration web table extraction and building a Mashup. J Comput Sci 7(2):129–142, ISSN 1549–3636
8. Zhai Y, Liu B (2005) Web data extraction based on partial tree alignment. In: Proceedings of WWW 2005, May 10–14 2005, Chiba, Japan. ACM 1-59593-046-9/05/0005
9. Zheleva E, Getoor L (2009) To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In: Proceedings of WWW 2009
10. He J, Chu WW, Liu Z (2006) Inferring privacy information from social networks. In: Proceedings of ISI, pp 154–165
11. Heatherly R, Kantarcioglu M,Thuraisingha B, Lindamood J (2009) Reventing private information inference attacks on social networks. Technical report UTDCS-03-09, University of Texas at Dallas
12. Tang C, Ross K, Saxena N, Chen R (2011) What's in a name: a study of names, gender inference, and gender behavior in facebook. DASFAA workshops 2011, pp 344–356
13. Nancy P, Geetha Ramani R (2012) Knowledge discovery (email harvesting, gender identification and prediction) in social network data (facebook 100 million URLs), Lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, 24–26 October, 2012, San Francisco, USA, pp 449–454
14. Popular baby names [Online] Available: http://www.ssa.gov.OACT/babynames
15. Facebook name list [Online] Available: http://sites.google.com/site/facebooknamelist/

Chapter 37
Human Gait Modeling and Statistical Registration for the Frontal View Gait Data with Application to the Normal/Abnormal Gait Analysis

Kosuke Okusa and Toshinari Kamakura

Abstract We study the problem of analyzing and classifying frontal view human gait data by registration and modeling on a video data. In this study, we suppose that frontal view gait data as a mixing of scale changing, human movements and speed changing parameter. Our gait model is based on human gait structure and temporal-spatial relations between camera and subject. To demonstrate the effectiveness of our method, we conducted two sets of experiments, assessing the proposed method in gait analysis for young/elderly person and abnormal gait detecting. In abnormal gait detecting experiment, we apply K-NN classifier, using the estimated parameters, to perform normal/abnormal gait detect, and present results from an experiment involving 120 subjects (young person), and 60 subjects (elderly person). As a result, our method shows high detection rate.

Keywords Abnormal gait detection · Frontal view gait data · Gait analysis · Human gait modeling · K-nearest neighbor classifier · parameter estimation · Scale registration · speed estimation · statistical registration

37.1 Introduction

We study the problem of analyzing and classifying frontal view gait video data. A study on the human gait analysis is very important in the field of the sports/health managements. For instance, gait analysis is one of the important method to detect the Alzheimer disease, infantile paralysis, and any other diseases [9, 16, 19].

K. Okusa (✉) · T. Kamakura
Department of Science and Engineering, Chuo University, 1128551 Tokyo, Japan
e-mail: k.okusa@me.com

T. Kamakura
e-mail: kamakura@indsys.chuo-u.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_37,

Gait analysis is mainly based on motion capture system and video data. The motion capture system can give the precise measurements of trajectories of moving objects, but it requires the laboratory environments and we cannot be used this system in the field study. On the other hand, the video camera is handy to observe the gait motion in the field study.

Gage [6] proposed brain paralysis gait analysis using gait video data. Kadaba et al. [8] discussed importance of lower limb in the human gait using gait video data too. Many gait analysis have recently analyzing using video analysis software (e.g. Dartfish, Contemplas, Silicon Coach). For example, Borel et al. [3] and Grunt et al. [7] proposed infantile paralysis gait analysis using lateral view gait video data.

On the other hand, from the standpoint of statistics, Olshen et al. [15] proposed the bootstrap estimation for confidence intervals of the functional data with application to the gait cycle data observed by the motion capture system. Soriano et al. [17] proposed the human recognition method based on the data matching techniques using the dynamic time warping of human silhouettes.

However, most studies have not focused on frontal view gait analysis, because such data has many restrictions on analysis based on the filming conditions. The video data filmed from the frontal view is difficult to analyze, because the subject getting close in to the camera, and data includes the scale-changing parameters [2, 10]. To cope with this, Okusa et al. [14] and Okusa and Kamakura [12] proposed a registration for scales of moving object using the method of nonlinear least squares.

Okusa et al. [14] and Okusa and Kamakura [12] did not focus on the human leg swing. Okusa and Kamakura [11] focus on the gait analysis using arm and leg swing model estimated parameters. However, their work was not enough to describe about further details of human gait modeling. In this chapter, we describe further details about Okusa and Kamakura [11] models and its applications.

We suppose that frontal view gait data as a mixing of scale changing, human movements and speed changing parameter. We estimate these parameters using the statistical registration and modeling on a video data. Our gait model is based on the human gait structure and temporal-spatial relations between camera and human.

To demonstrate the effectiveness of our method, we conducted two sets of experiments, assessing the proposed method in gait analysis for young/elderly person and abnormal gait detection.

In young/elderly gait analysis experiment, we analyze young 120 subjects and elderly 60 subjects (normal gait: 46, abnormal gait: 14). We discuss the normal and abnormal gait features using the estimated parameters from proposed model. In abnormal gait detection experiment, we apply a K-nearest-neighbor (K-NN) classifier, using the estimated parameters, to perform abnormal gait detection, and present results from an experiment involving 180 subjects from young/elderly gait analysis. As a result, our method shows high detection rate.

37.2 Frontal View Gait Data

In this section, we describe an overview of frontal view gait data. Many of gait analysis using lateral view gait data, because lateral view gait is easy to detect the human gait features. However, in a corridor like structure, the subject is approaching a camera. Such case is difficult observe lateral view gait.

In a lateral view gait, at least two cycles or four steps are needed. For more robust estimation of the period of walking, about 8 m is recommended. To capture this movement, the camera distance required is about 9 m. Practically, having such a wide space is difficult. On the other hand, frontal view gait video is easy to observe 8 m (or more) gait steps [10].

Figure 37.1 is an example of frontal view gait data recorded by Fig. 37.2 situation. Figure 37.1 illustrates difficulty of frontal view gait analysis. Even if subject do the same motion with the same timing, frontal view gait data is included scale changing components. Figure 37.3 shows subject width time-series behavior of frontal view gait data. This figure illustrates frontal view gait data contains many time-series components.

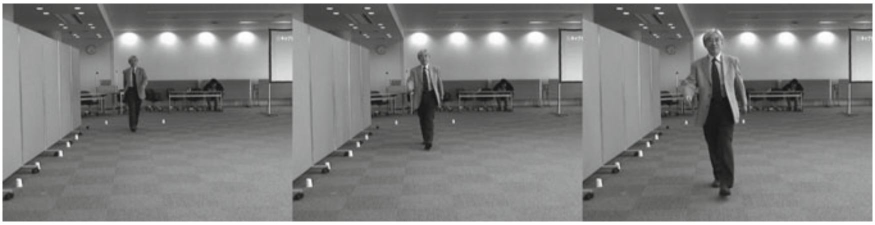
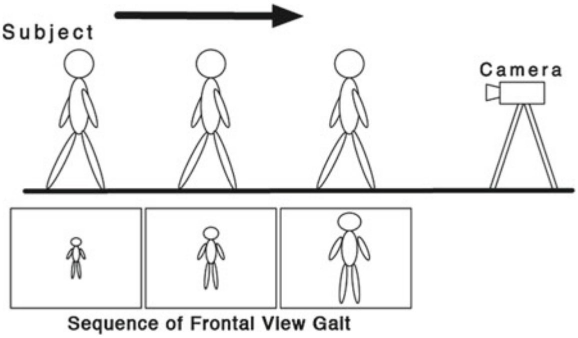


Fig. 37.1 *Frontal view* gait data

Fig. 37.2 Filming situation of *frontal view* gait data: In this situation, subject's apparent scale is changing because subject getting close to the camera



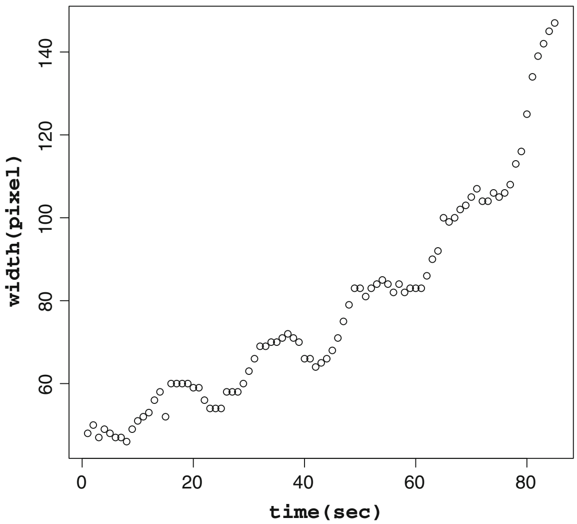


Fig. 37.3 Time-series behavior of *frontal view* subject width: x-axis and y-axis are time(sec) and width(pixel) respectively, and *white circle* means observed value. Data includes scale-changing, subject's movement, speed-changing components

37.3 Modeling of Frontal View Gait Data

37.3.1 Preprocessing

The raw video data is difficult to observe subject width and height time-series behavior because data contain background. We separate subject from background using inter-frame subtraction method (Eq. 37.1).

$$\varDelta^{(T)} = |I^{(T+1)} - I^{(T)}|, T = 1, ..., (n-1),$$
$$\varDelta^{(T)}(p,q) = \begin{cases} 1 & (\varDelta^{(T)}(p,q) > 0) \\ 0 & (\text{Otherwise}). \end{cases} \qquad (37.1)$$

Here, $\varDelta^{(T)}$ is an inter-frame subtraction image, $I^{(T)}$ is grey scaled video data image at frame T, (p,q) is the pixel coordinate.

37.3.1.1 Subject Width/Height Calculation

Inter-frame subtraction method can separate the subject and background. However, it is difficult to measure the time-series behavior of the subject width and height. In this section, we describe the subject width and height calculation method using inter-frame subtraction data.

Let us suppose that inter-frame subtraction image is binary matrix. We can measure the subject height and width by integration calculation of row and column at each frame. In this study, we focus on the human gait arm and leg swing of the frontal view gait. We assume that subject width and height time-series behavior consist of the arm and leg swing behavior.

37.3.2 Relationship Between Camera and Subject

Figure 37.4 shows a relationship between camera and subject. Width and height modeling has same structure. In this section, we describe the subject width modeling. We can assume simple camera structure. We consider the virtual screen exists between observation point and subject, and we define subject width on the virtual screen x_i at i-th frame($i = 1, ..., n$).

Here we define z_i, z_j as distance between observation point and subject at i-th, j-th frame, z_s as distance between observation point and virtual screen, $\theta_{x_{i1}}$, $\theta_{x_{i2}}$ as subject angle of view from observation point at i-th frame, d as distance between observation point and 1st frame, v_i as subject speed at i-th frame. Okusa et al. [14] defined the subject length L was constant. We assume that L has the time-series behavior and we define L_i is the subject length at i-th frame.

x_i at i-th frame depends on $\theta_{x_{i1}}$, $\theta_{x_{i2}}$ as shown in Fig. 37.4.

$$x_i = z_s(\tan\theta_{x_{i1}} + \tan\theta_{x_{i2}}). \tag{37.2}$$

Similarly, the subject length at i-th frame is

$$L_{x_i} = z_i(\tan\theta_{x_{i1}} + \tan\theta_{x_{i2}}). \tag{37.3}$$

From Eqs. (37.2) and (37.3), ratio between x_n and x_i is

$$\frac{x_n}{x_i} = \frac{L_{x_n} z_i}{L_{x_i} z_n} \tag{37.4}$$

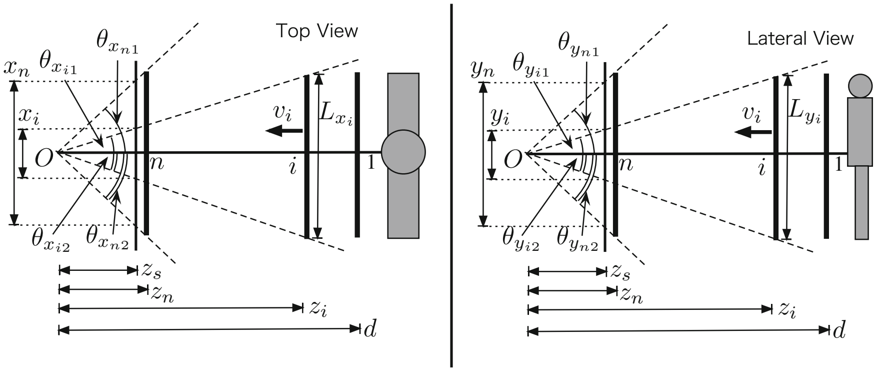


Fig. 37.4 Relationship between camera and subject

Frame interval is equally-spaced (15 fps). Okusa et al. [14] assumed the average speed is constant. We can assume that average speed from i-th frame is $(n-i) = (z_i - z_n)/\bar{v}$, therefore z_i is $z_i = z_n - \bar{v}(n-i)$. We substitute z_i to Eq. (37.4)

$$x_i = \frac{M_{x_i}\gamma}{\gamma + (n-i)} x_n + \epsilon_i, \tag{37.5}$$

where γ is $z_n/\bar{v}$, M_{x_i} is L_{x_i}/L_{x_n}, ϵ_i is noise. From Eq. (37.5), predicted value $\hat{x}_i^{(n)}$ is registration from i-th frame's scale to n-th frame's scale

$$\hat{x}_i^{(n)} = \frac{\gamma + (n-i)}{M_{x_i}\gamma} x_i. \tag{37.6}$$

Similarly, we can define subject height as

$$y_i = \frac{M_{y_i}\gamma}{\gamma + (n-i)} y_n + \epsilon_i, \tag{37.7}$$

where M_{y_i} is L_{y_i}/L_{y_n}.

Next, we discuss the scale changing, human movement, and speed changing parameter estimation model.

37.3.3 Scale Changing Parameter Estimation

From Eq. (37.5), scale parameter is γ. Solve Eq. (37.5) for γ shows

$$\gamma = \frac{x_i(n-i)}{x_i - M_{x_i} x_n}. \tag{37.8}$$

Here γ is the ungaugeable parameter, and we estimate it using nonlinear least squares method

$$S(\gamma, M_{x_i}) = \sum_{i=1}^{n} \left\{ x_i - \frac{M_{x_i}\gamma}{\gamma + (n-i)} x_n \right\}^2. \tag{37.9}$$

37.3.4 Human Movement Parameter Estimation

M_{x_i} and M_{y_i} are movement model of the subject. If the subject is the rigid body, movement model M_{x_i} and M_{y_i} are constant. Meanwhile, human gait is not a constant. M_{x_i} and M_{y_i} needs the movement model because the subject body is moving wildly.

37.3.4.1 Human Gait Modeling: Arm Swing

Collins et al. [5] has reported that arm swing is an very important role in the gait motion based on the simple gait model. We consider the human gait modeling based on Collins et al. [5] model (see Fig. 37.5).

It seems reasonable to think that arm is single pendulum. Collins et al. [5] model assumed the arm swing is move to anteroposterior direction. Our model, on the other hand, can assume that arm swing move to an oblique direction (Fig. 37.6).

Figure 37.6s model has an ungaugeable area. Our method's width/height calculation is based on integration calculation of row and column at each frame. If the arm move to inside body area, arm length is ungaugeable. Arm swing model is

$$\begin{aligned} x_i &= \frac{\left(\frac{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,i)}{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,n)} + s\right)\gamma}{\gamma + (n-i)} x_n + \epsilon_i \\ W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,i) &= P_1\tau(fi + Q_1, g_1) + P_2\tau(fi + Q_2, g_2) \\ \tau(\theta, g) &= \begin{cases} \sin(\theta) + g & (\sin(\theta) + g > 0) \\ 0 & (\text{Otherwise}) \end{cases} \end{aligned} \tag{37.10}$$

where $P_1 = a_1\cos(\psi)$ and $P_2 = a_2\cos(\psi)$. $P_1\tau(fi + Q_1, g_1)$ and $P_2\tau(fi + Q_2, g_2)$ are right and left arm model respectively. From Eq. (37.10), we estimate each gait parameter using nonlinear least squares method.

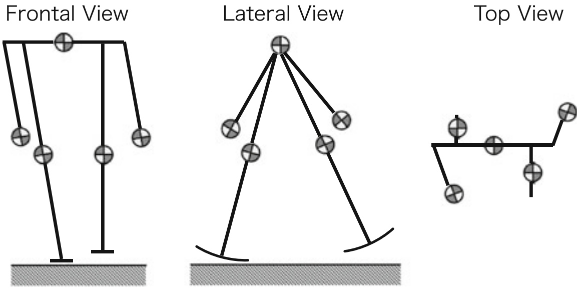


Fig. 37.5 Gait model

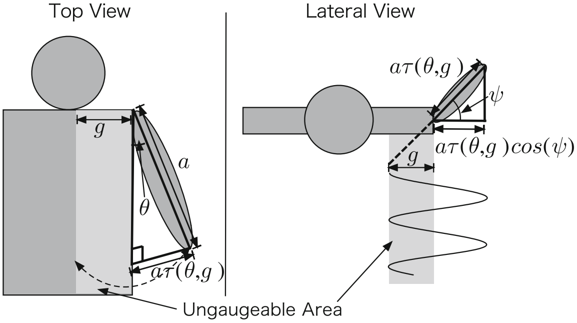


Fig. 37.6 Arm swing model: This gait model is based on the Collins et al. [5] model. Collins et al. [5] model assumed the arm swing is move to *anteroposterior direction*. Our model, on the other hand, can assume that arm swing move to an *oblique direction*

$$
\begin{aligned}
&S(\gamma,P_1,P_2,Q_1,Q_2,g_1,g_2,f,s) = \\
&\sum_{i=1}^{n} \left\{ x_i - \frac{\left(\frac{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,i)}{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,n)} + s \right)\gamma}{\gamma + (n-i)} x_n \right\}^2 .
\end{aligned}
\tag{37.11}
$$

Here, f is gait cycle frequency, s is adjustment parameter, P_1, P_2 are arm swing amplitude parameters, Q_1, Q_2 are arm phase parameters, and g_1, g_2 are ungaugeable area parameters.

37.3.4.2 Human Gait Modeling: Leg Swing

The leg swing modeling is simpler than arm swing model because the leg model does not have a ungaugeable area. Okusa et al. [14] and Okusa and Kamakura [12] does not consider the leg swing. It seems reasonable to think like arm swing that leg swing is single pendulum (Fig. 37.7).

Leg swing model is

$$
\begin{aligned}
y_i &= \frac{\left(\frac{H(b_1,Q_3,f,i)}{H(b_1,Q_3,f,n)} + s \right)\gamma}{\gamma + (n-i)} y_n + \epsilon_i \\
H(b_1,Q_3,f,i) &= b_1 \cos(fi + Q_3).
\end{aligned}
\tag{37.12}
$$

Here b_1 is leg swing amplitude parameter, and Q_3 is leg phase parameter.

37.3.5 Speed Changing Parameter Estimation

Frontal view video data is difficult to see the subject's speed. If our gait model is correct, observed value x_i and y_i is same as the fitted value of gait model at point ℓ_i. Previous model's ℓ_i assumes equally spaced ($\ell_i = i = 1, \ldots, n$). We estimate ℓ_{x_i} and

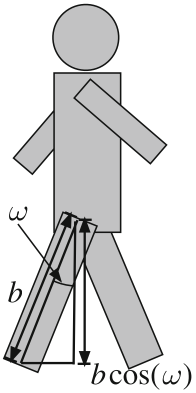


Fig. 37.7 Leg swing model: The leg swing modeling is simpler than arm swing model because the leg model does not have a ungaugeable area

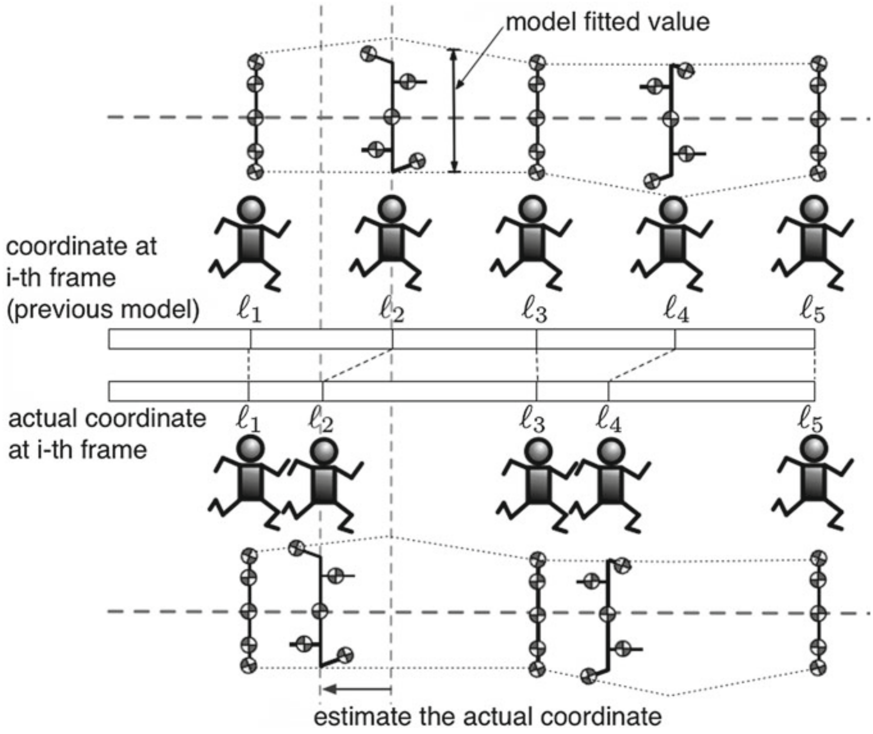


Fig. 37.8 Virtual space coordinate estimation

ℓ_{y_i} value for minimize the observed value and model fitted value at ℓ_i. We can define estimated value ℓ_{x_i} and ℓ_{y_i} as a virtual space coordinate at i-th frame (Fig. 37.8).

Equations. 37.5 and 37.7 with the coordinate estimation shows

$$\begin{aligned} x_i &= \frac{M_{x_i}\gamma}{\gamma + (n - \ell_{x_i})} x_n + \epsilon_i \\ y_i &= \frac{M_{y_i}\gamma}{\gamma + (n - \ell_{y_i})} y_n + \epsilon_i. \end{aligned} \tag{37.13}$$

Here, $\ell_{x_i}, ..., \ell_{x_n}$ and $\ell_{y_i}, ..., \ell_{y_n}$ are virtual space coordinate parameters of width and height respectively. From Eq. 37.13, arm swing and leg swing model with the coordinate estimation shows Eqs. 37.14 and 37.15.

$$
\begin{aligned}
x_i &= \frac{\left(\frac{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,\ell_{x_i})}{W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,\ell_{x_n})}+s\right)\gamma}{\gamma+(n-\ell_{x_i})}x_n+\epsilon_i \\
W(P_1,P_2,Q_1,Q_2,g_1,g_2,f,\ell_{x_i}) &= P_1\tau(f\ell_{x_i}+Q_1,g_1)+P_2\tau(f\ell_{x_i}+Q_2,g_2) \\
\tau(\theta,g) &= \begin{cases} \sin(\theta)+g & (\sin(\theta)+g>0) \\ 0 & (\text{Otherwise}). \end{cases}
\end{aligned}
\tag{37.14}
$$

$$
y_i = \frac{\left(\frac{H(b_1,Q_3,f,\ell_{y_i})}{H(b_1,Q_3,f,\ell_{y_n})}+s\right)\gamma}{\gamma+(n-\ell_{y_i})}y_n+\epsilon_i \tag{37.15}
$$

$$
H(b_1,Q_3,f,\ell_{y_i}) = b_1\cos(f\ell_{y_i}+Q_3).
$$

We suppose that virtual space coordinate of subject is $\hat{\ell}_i = (\hat{\ell}_{x_i}+\hat{\ell}_{y_i})/2$. Then, we can assume that subjects speed is 1st order difference of $\hat{\ell}_i$, and acceleration is 2nd order difference of $\hat{\ell}_i$.

We estimate these models parameters using Okusa and Kamakura [13] method. This method is very stable and very fast to estimate these parameters.

37.3.6 Gait Parameter Estimation

In this section, we discuss the gait parameter estimation.

Figure 37.9 is plot of the subject width (left side) and height (right side) time-series behavior. In Fig. 37.9 (left side), white circle is observed value, dotted line represent fitted value of Eq. 37.5 (scale variant estimation model), continuous line represent fitted value of Eq. 37.10 (scale variant + arm movements estimation

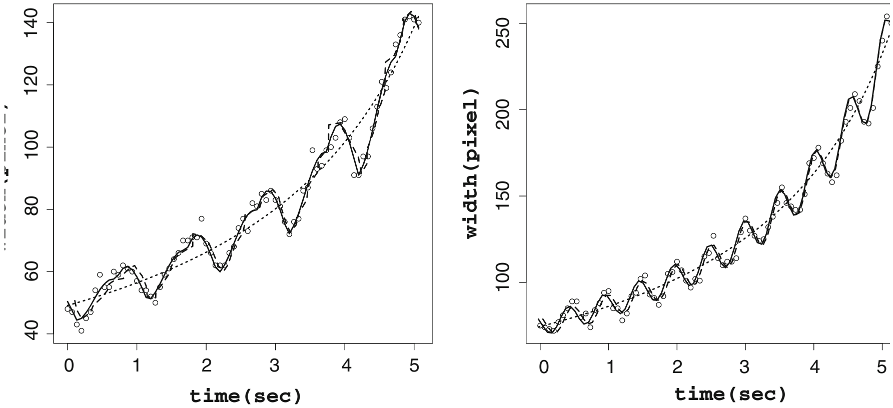


Fig. 37.9 Model fitting: width (*left side*), height (*right side*)

Table 37.1 RSS, AIC, cAIC of width data

Method	RSS	AIC	cAIC
1st order regression	6,834.98	569.94	569.99
2nd order regression	3,808.77	526.91	527.08
Eq. 37.5 model	3,077.40	508.50	510.66
Eq. 37.10 model	425.63	372.17	374.86
Eq. 37.14 model	189.46	463.84	−1,032.56

Table 37.2 RSS, AIC, cAIC of height data

Method	RSS	AIC	cAIC
1st order regression	21,502.96	665.65	665.70
2nd order regression	7,149.68	581.77	581.93
Eq. 37.7 model	5,248.84	555.66	557.82
Eq. 37.12 model	631.11	406.43	409.08
Eq. 37.15 model	401.15	519.09	−1804.91

model), dashed line represent fitted value of Eq. 37.14 (scale variant + arm movements + speed variant estimation model).

Similarly, in Fig. 37.9 (right side), white circle is observed value, dotted line represent fitted value of Eq. 37.7 (scale variant estimation model), continuous line represent fitted value of Eq. 37.12 (scale variant + leg movements estimation model), dashed line represent fitted value of Eq. 37.15 (scale variant + leg movements + speed variant estimation model).

Table 37.1 is Residual Sum of Squares (RSS), Akaike Information Criterion (AIC) [1] and Consistent Akaike's Information Criterion (cAIC) [18] value of first order regression, second order regression, Eqs. 37.5, 37.10 and 37.14 models of width data. Table 37.2 is RSS, AIC and cAIC value of first order regression, second order regression, Eqs. 37.7, 37.12 and 37.15 models of height data.

In Tables 37.1 and 37.2, most minimal AIC model is Eqs. 37.10 and 37.12. Meanwhile, most minimal RSS and cAIC model are Eqs. 37.14 and 37.15. Burnham and Anderson [4] strongly recommend using cAIC, rather than AIC, if number of data n is small or number of parameters k is large. Since cAIC converges to AIC as n gets large, cAIC generally should be employed regardless. Therefore, we select the Eqs. 37.14 and 37.15 model. Figure 37.9 and Tables 37.1 and 37.2 illustrates our method has a good performance.

In next section, we discuss the effectiveness of our method.

37.4 Experimental Details and Results

To demonstrate the effectiveness of our method, we conducted two sets of experiments, assessing the proposed method in gait analysis for young/elderly person and abnormal gait detection. We use SONY DCR-TRV70K camera. Frame rate of video data is 15 fps and resolution is 640 × 480.

In this chapter, we focus on $\hat{\gamma}$ (speed parameter), $(\hat{P}_1 + \hat{P}_2)/2$ (width amplitude parameter), and $\hat{b}_1$ (height amplitude parameter).

37.4.1 Gait Analysis: Young Person

In this experiment we took movie of 120 subjects walking video data from frontal view [10 steps, Male: 96 (average height: 173.24 cm, sd: 5.64 cm), Female: 24 (average height: 156.25 cm, sd: 3.96 cm)] and apply to our proposed method for the gait analysis.

Figure 37.10 is width amplitude versus speed (left side) and height amplitude versus speed (right side). The important point to note is that speed parameter $\hat{\gamma}$ is $z_n/\bar{v}$. If the subject walking fast, speed parameter $\hat{\gamma}$ is small. From Fig. 37.10, width amplitude versus speed and height amplitude versus speed have a nonlinear relationship. This results means, if the subject's arm swing and leg swing moving strongly, subject's walking speed is fast. Moreover, these relationships are "nonlinear".

37.4.2 Gait Analysis: Elderly Person

In this experiment we took movie of 60 subjects walking video data from frontal view (10 steps / average age: 76.97· sd: 4.16 / abnormal gait subjects; average age: 77.56 · sd: 4.35 / normal gail subjects; average age: 75.37· sd: 3.18) and apply to our proposed method for the gait analysis.

Figure 37.11 is width amplitude versus speed (left side) and height amplitude versus speed (right side). Black and white circle means normal and abnormal gait

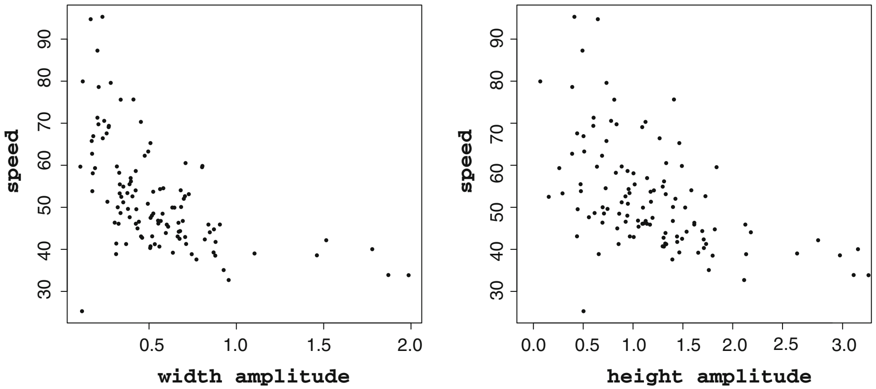


Fig. 37.10 Young subject: width amplitude versus speed (*left side*), height amplitude versus speed (*left side*)

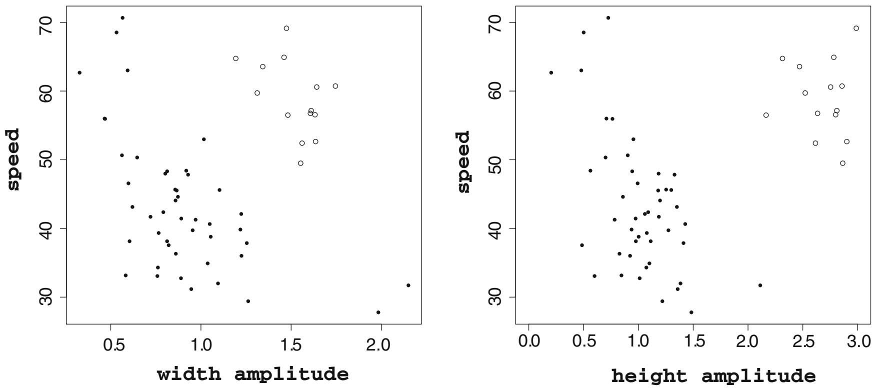


Fig. 37.11 Elderly subject: width amplitude versus speed (*left side*), height amplitude versus speed (*left side*)

Table 37.3 Normal/Abnormal gait average detection rate (%)

	Normal gait	Abnormal gait
Normal gait	98.2	1.8
Abnormal gait	0	100

subjects respectively. From Fig. 37.11, normal gait subject width amplitude versus speed and height amplitude versus speed have a nonlinear relationship like a young person. However, on the other hand, abnormal gait subject estimated parameters does not have nonlinear relationship. These abnormal gait parameters clustered in different place from normal gait subjects. The result leads to our presumption that the abnormal gait subject trying to moving fast, but this effort is not effective to moving speed.

37.4.3 Abnormal Gait Detection

In this section, we apply K-NN classifier (K = 3), using the estimated parameters, to perform normal/abnormal gait detect, and present results from an experiment involving 120 subjects (young person), and 60 subjects (elderly person).

To evaluate our estimated parameters, we apply these parameters to leave-one-out cross-validation test. Table 37.3 is average detection rate of normal/abnormal gait. Table 37.3 shows our estimated parameters may be used for the normal/abnormal gait detection.

37.5 Conclusions

In this article, focusing on the human gait cycles, we consider the human gait modeling based on simple gait structure. We estimate the parameters of the human gait cycles using the method of nonlinear least squares.

We also show that estimated parameters may be used for the human gait analysis and abnormal gait detection. Experimental results verify that our model can estimate the various parameters and these parameters are good feature values of the human gait. We plan to implement this scheme for the sports analysis of the long-distance runner.

References

1. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. International symposium on information theory, pp 267–281
2. Barnich O, Droogenbroeck MV (2009) Frontal-view gait recognition by intra-and inter-frame rectangle size distribution. Pattern Recogn Lett 30:893–901
3. Borel S, Schneider P, Newman CJ (2011) Video analysis software increases the interrater reliability of video gait assessments in children with cerebral palsy. Gait Posture 33(4):727–729
4. Burnham KP, Anderson DR (2010) Model selection and multi-model inference: a practical information-theoretic approach. Springer, New York
5. Collins SH, Adamczyk PG, Kuo AD (2009) Dynamic arm swinging in human walking. Proc R Soc B: Biol Sci 276(1673):3679–3688
6. Gage JR (1982) Gait analysis for decision-making in cerebral palsy. Bull Hosp Jt Dis Orthop Inst 43(2):147–163
7. Grunt S, van Kampen PJ, Krogt MM, Brehm MA, Doorenbosch CAM, Becher JG (2010) Reproducibility and validity of video screen measurements of gait in children with spastic cerebral palsy. Gait Posture 31(4):489–494
8. Kadaba MP, Ramakrishnan HK, Wootten ME (1990) Measurement of lower extremity kinematics during level walking. J Orthop Res 8(3):383–392
9. Kirtley C (2006) Clinical gait analysis: theory and practice, 1e, 1 edn. Churchill Livingstone
10. Lee TKM, Belkhatir M, Lee PA (2008) Fronto-normal gait incorporating accurate practical looming compensation. Pattern Recogn
11. Okusa K, Kamakura T (2012) Normal/Abnormal gait analysis based on the statistical registration and modeling of the frontal view gait data. Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science (2012) WCECS 2012, 24–26 Oct 2012. USA, San Francisco, pp 443–448
12. Okusa K, Kamakura T (2012) A statistical registration of scale changing and moving objects with application to the human gait analysis. (in Japanese). Bull Jpn Soc Comput Statist 24(2)
13. Okusa K, Kamakura T (2012) Statistical registration and modeling of frontal view gait data with application to the human recognition. In: International conference on computer statistics (COMPSTAT 2012), pp 677–688
14. Okusa K, Kamakura T, Murakami H (2011) A statistical registration of scales of moving objects with application to walking data (in Japanese). Bull Jpn Soc Comput Statist 23(2):94–111
15. Olshen RA, Biden EN, Wyatt MP, Sutherland DH (1989) Gait analysis and the bootstrap. Ann Statist, pp 1419–1440

16. Perry J, Burnfield J (2010) Gait analysis: normal and pathological function. Slack Incorporated
17. Soriano M, Araullo A, Saloma C (2004) Curve spreads-a biometric from front-view gait video. Pattern Recogn Lett 25(14):1595–1602
18. Sugiura N (1978) Further analysts of the data by Akaike's information criterion and the finite corrections. Commun Statist Theory Methods
19. Whittle MW (1991) Gait analysis: an introduction. Butterworth-Heinemann

Chapter 38
Statistical Recognition of Aspiration Presence

Shuhei Inui, Kosuke Okusa, Kurato Maeno and Toshinari Kamakura

Abstract A study on the healthcare application is very important for the solitary death in aging society. Many previous works had been proposed a detection method of aspiration using the non-contact radar. But the works are only in subjects with sitting in a chair. We consider that user falls down in the state when he happen abnormal situation as daily life. In this study, we focus on the detection of "aspiration" or "apnea" for the lying position, because the final decision of the life or death is aspiration. As initial stage of the system, we propose the recognition method for the presence of aspiration with lying position under the low-disturbance environment from microwave Doppler signals by using support vector machine (SVM).

Keywords Aspiration modeling · Healthcare · Microwave Doppler radar · Monitoring system · Signal processing · Support vector machine

S. Inui (✉)
Graduate School of Science and Engineering, Chuo University, 1-13-27 Kasuga,
Bunkyo-ku, Tokyo 112-8551 , Japan
e-mail: s.inui.1222@gmail.com

K. Okusa
Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku,
Tokyo 112-8551, Japan
e-mail: k.okusa@me.com

K. Maeno
Corporate Research and Development Center, Oki Electric Industry Co., Ltd, 1-16-8 Chuo,
Warabi-shi, Saitama 335-8510, Japan
e-mail: maeno284@oki.com

T. Kamakura
Faculty of Science and Engineering, Chuo University, 1-13-27 Kasuga, Bunkyo-ku, Tokyo
112-8551, Japan
e-mail: kamakura@indsys.chuo-u.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_38,

38.1 Introduction

Recently, a study on the healthcare application is very important for the solitary death in aging society. The non-contact radars attract an attention because the system requires that daily life of user does not interfere. Many of these radars (e.g. infrared radar, sound radar, Doppler radar) have used in the system, especially the microwave Doppler radar has the advantage against the noise, light and temperature than another radars. Therefore, this feature is considered to be suitable for application to the system. However, the radar receives all environment movements in this area. The radar is not robust under the disturbance environment. Most studies have not focused on the radar.

As perspective of monitoring for the elderly, we focus on the detection of "aspiration" or "apnea" for the lying position, because we think the final decision of the life or death is aspiration. As initial stage of the system, this study proposes the recognition method for the presence of aspiration with lying position under the low-disturbance environment from microwave Doppler signals by using support vector machine (SVM). If the presence of aspiration with lying position can be recognized, it is expected to apply to the detection of apnea syndrome and solitary death.

We describe the microwave Doppler radar system in Sect. 38.2, we review other studies in Sect. 38.3. We define the proposed method in Sect. 38.4, we explain our experimental condition of proposed method in Sect. 38.5 and we discuss result of analysis. Finally, we discuss to the conclusion in Sect. 38.7.

38.2 Microwave Doppler Radar

We describe the microwave Doppler radar system. The radar irradiates the microwave to the target and let F_a be the frequency of transmitted wave. The wave hits the target, and let F_b be the frequency of reflected wave toward the radar. The frequency of Doppler subtracts F_a from F_b. The radar outputs the electric signal according to it.

In this study, the radar uses the IPS-154 manufactured by Innocent Co., Ltd, and the A/D converter uses the USB2.0 compatible analog output terminal manufactured by Contec Co., Ltd. (Fig. 38.1).

The radar is classified into two types (dual type and single type) which differ to the output wave. It derives two outputs V_I and V_Q, which have a quadrature phase relationship, that is to say their phases are 90° different from each other. If we do not consider the noise, both V_I and V_Q are shown below.

$$V_I = A_1 \sin\left(\frac{4\pi R_1}{\lambda}\right) \tag{38.1}$$

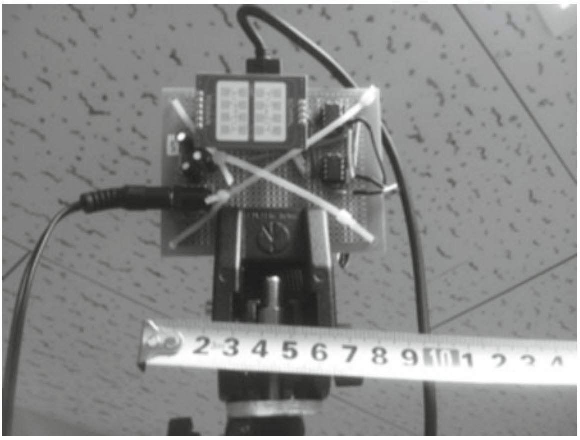


Fig. 38.1 Microwave Doppler radar (IPS-154)

$$V_Q = A_2 \cos\left(\frac{4\pi R_2}{\lambda} + \phi\right) \tag{38.2}$$

where A_1 and A_2 are amplitudes, λ is the wave length, R is the distance between the radar and target and ϕ is the initial phase.

It follows from Eqs. (38.1) and (38.2) that the phase change $\Delta\phi$ is proportional to the range change between the target and the radar ΔR.

$$\Delta_\phi = \frac{4\pi\Delta R}{\lambda} \tag{38.3}$$

Instantaneous amplitude A and phase difference ϕ_t are shown below.

$$A = \sqrt{V_I^2 + V_Q^2} \tag{38.4}$$

$$\phi_t = \tan^{-1}\frac{V_Q}{V_I} \tag{38.5}$$

38.3 Related Work

We review the related works of this study. Aoki et al. [1] focused on the behavior patterns of solitude senior using pyroelectric radar, then they proposed the detection method of irregular states. Kubo et al. [2] proposed a human activity recognition method based on the Doppler radar using three binary classifiers (least squares, SVM and AdaBoost) approach. They detected the move (the target is changing his/her position or pose), resp (the target sits still and is aspirating) and hold (the target sits still and holds his/her aspiration) movement.

Kubo et al. [3, 4] proposed the aspiration wave estimation method using the microwave Doppler radar. They introduced a criterion in evaluating the phase estimation. Then, they proposed five methods (offset estimation, mean, least squares method, estimation based on raster images, estimation based on Monte Carlo method) to estimate the signal phase and compared their performances by computer simulation and experiment. Lien et al. [6] proposed the aspiration and heartbeat detection method based on the millimeter-wave Doppler radar system using root-MUSIC method. Naoi et al. [7] proposed the heart beat detection method based on the microwave Doppler radar using time-difference approach.

Petrochilos et al. [8] focused on distinguishing aspiration from heartbeat, then they verified method of RACMA and ICA. Sekine et al. [9] proposed a human activity (e.g. shaking hands, walking, etc...) recognition algorithm based on the microwave Doppler radar using Support Vector Machine (SVM). Tanigawa et al. [10] proposed a human chewing detection method based on the microwave Doppler radar using wavelet transform and auto-correlation coefficient. Zhou et al. [11] proposed the heartbeat wave model under the multi Doppler radar environment. However, their method is difficult to set the initial value.

38.4 Proposed Method

We define the proposed method.In this study, we discuss recognizing the presence of aspiration with lying position.

Firstly, the proposed method removes high frequency component. Generally, the aspiration component is distributed in 0.3 Hz area, and the heartbeat component is distributed in 1–1.2 Hz area. Human activity is distributed in low frequency. We filter the received signal with a low-pass filter (see Fig. 38.2), so a pass band sets up 0–2 Hz. Figure 38.2s x-axis means time index, and y-axis means Voltage (V) of microwave Doppler signals. Above figure of Fig. 38.2 is observed data, and below figure of Fig. 38.2 is low-pas filtered data. Many of low-pass filter method had been reported. In this study, we apply the Fast Fourier Transform (FFT) based on the method.

Secondly, Eq. (38.4) produces instantaneous amplitude of I-Q signal from low-pass filtered data.

Thirdly, we calculate feature quantities for the SVM. We are focusing mean, variance, maximum value, minimum value, skewness and kurtosis of instantaneous amplitude of I-Q signal, then these values consider feature quantities for the SVM. Window size is set to 1,000 samples, and window sift size is set to 1 sample when we calculate feature quantities for the SVM.

Fourthly, Eq. (38.7) estimates amplitude, frequency, phase of instantaneous amplitude of I-Q signal. Estimation result of each parameter considers feature quantities for the SVM. Estimation model is shown below.

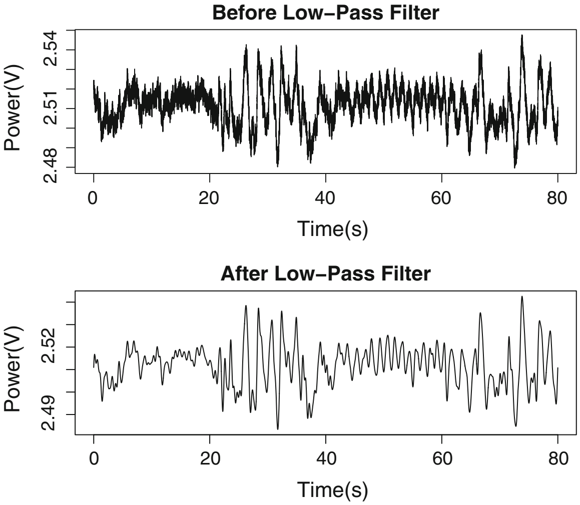


Fig. 38.2 Comparision before and after the low-pass filter

$$y(x_i) = A\sin\left(2\pi\omega x_i \frac{1}{f} + \phi_1\right) + A\cos\left(2\pi\omega x_i \frac{1}{f} + \phi_2\right) + \varepsilon_i \tag{38.6}$$

where A is amplitude, ω is angular frequency, f is sampling frequency and ϕ_1, ϕ_2 are phases. Let x_i be data index runs from $i = 1, \ldots, n$.

We minimize the Eq. (38.7), and estimate each parameter. We use the R (http://www.r-project.org/) for estimation of each parameter.

$$S(A, \omega, \phi_1, \phi_2) = \sum_{i=1}^{n} \left\{ y_i - \left(A\sin\left(2\pi\omega x_i \frac{1}{f} + \phi_1\right) + A\cos\left(2\pi\omega x_i \frac{1}{f} + \phi_2\right)\right)\right\}^2 \tag{38.7}$$

The initial values are set as follow: A is half value of the range of instantaneous amplitude of I-Q signal, ω is 0.3 (aspiration frequency) and ϕ_1, ϕ_2 are 0 (We assume that these phases does not exist). Window size is set to 1,000 samples, and window sift size is set to 1 sample when we estimate each parameters. Figure 38.3 shows result of estimation. Figure 38.3s x-axis means time index, and y-axis means Voltage (V) of instantaneous amplitude of I-Q signal.

Finally, we detect the two situations of "aspiration" and "apnea" using 2-class SVM (Gaussian Kernel). We use 10 feature quantities for the SVM like; mean, variance, maximum, minimum, skewness, kurtosis, A, f, ϕ_1 and ϕ_2.

Figure 38.4 shows flowchart of proposed method.

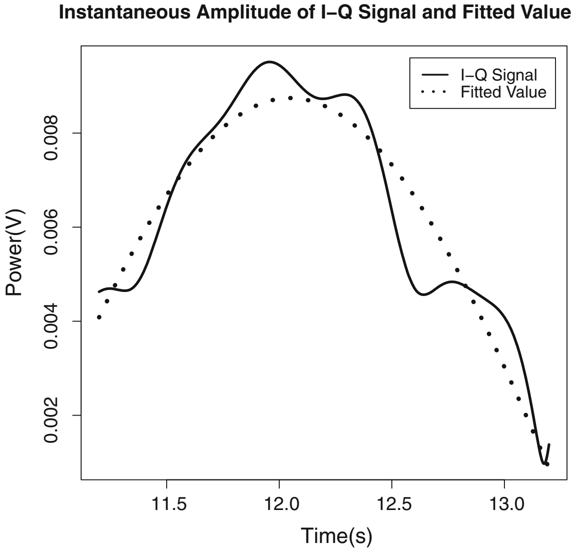


Fig. 38.3 Instantaneous amplitude of I-Q signal and fitted value

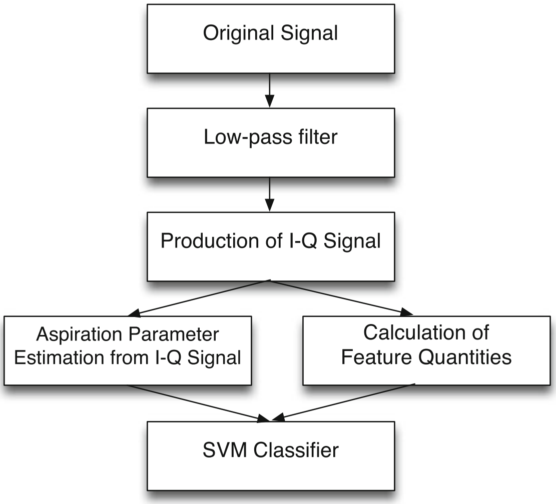

Original Signal
Low-pass filter
Production of I-Q Signal
Aspiration Parameter Estimation from I-Q Signal
Calculation of Feature Quantities
SVM Classifier

Fig. 38.4 Flowchart of proposed method

38.4.1 Feature Quantity for the SVM

Method of calculating the feature quantity for the SVM is shown below. $x_1, x_2, \ldots .x_N$ stand for N pieces of data, then $x_1 \leq x_2 \leq \ldots \leq x_N$ stand for N pieces of data arranged in ascending order.

Mean(μ)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Variance(σ^2)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Maximum value

$$\text{Maximum value} = x_N$$

Minimum value

$$\text{Minimum value} = x_1$$

Skewness(γ_1)

$$\gamma_1 = \frac{1}{N\sigma^3}\sum_{i=1}^{N}(x_i - \mu)^3$$

kurtosis(γ_2)

$$\gamma_2 = \frac{1}{N\sigma^4}\sum_{i=1}^{N}(x_i - \mu)^4 - 3$$

38.4.2 SVM

We introduce SVM based on Ref. [4]. SVM uses the linear threshold element as the simplest neuron model (see Fig. 38.5), constitutes the pattern of the two-class classifier. For the input feature vector, the linear threshold element calculates the binary output value by the discriminant function Eq. (38.8),

$$y = \text{ sign } (w^t x - h) \tag{38.8}$$

where w is parameter correspond to synaptic weight, h is threshold value, sign (u) is sign function.

If u is positive value, sign (u) is 1. Meanwhile, if u is negative value, sign (u) is -1. If the dot product of the input vector and the synaptic weight exceed the threshold value, then the model outputs 1. Otherwise, the model outputs -1.

Two-class defines C1 and C2, each class label is 1 and -1. N number of feature vectors $x_1, \ldots, x_N$ define training sample. Correct class label for each sample define

Fig. 38.5 Neuron model

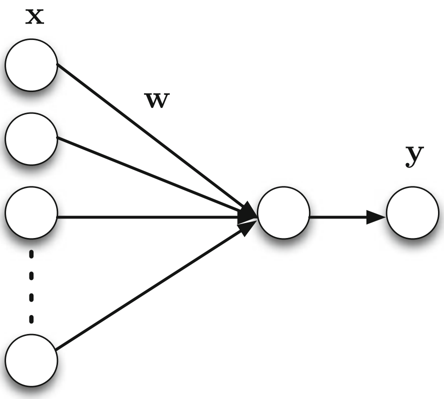
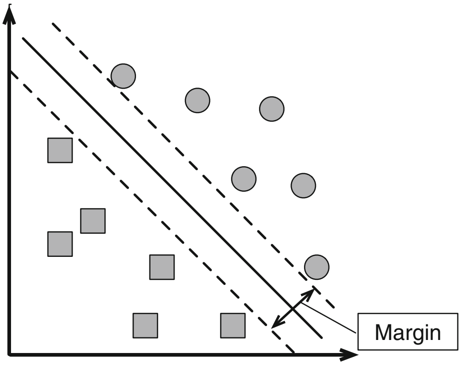


Fig. 38.6 Margin in the SVM



$t_1, \ldots, t_N$. Training sample assumes linearly separable. Distance of the identification plane and the closest training sample refer to margin (see Fig. 38.6). SVM purposes that the identification plane draws the maximum of margin.

If training sample is linearly separable,

$$t_i(w^t x_i - h)1 \geq 1, \quad i = 1, \ldots, N \tag{38.9}$$

There is the parameter to fulfill Eq. (38.9). Distance of the identification plane and the hyperplane $(\mathrm{H1} : w^t x - h = 1,\ \mathrm{H\,2} : w^t x - h = -1)$ becomes $\frac{1}{\|x\|}$. The problem of parameter w and h for Maximum of margin are equal to the problem of parameter for minimum of objective function Eq. (38.11) under the limiting condition Eq. (38.10).

$$t_i(w^t x_i - h)1 \geq 1, i = 1, \ldots, N \tag{38.10}$$

$$L(w) = \frac{1}{2}\|w\|^2 \tag{38.11}$$

The optimization problem is known to the quadratic programming problem a field of the mathematical programming, a variety of numerical methods have been proposed.

38.5 Outline of the Experiment

We explain our experimental condition of proposed method. In this study, we focus on the lying position subject's aspiration.

Our experimental condition assumed a eight-mat room. Space of square (3.6 by 3.6 m) reserved, then angle of the square made a paul (see Fig. 38.7). The radar was seated at the place where height was 2.3 m and 40° angle to the ground. Subject lain position at center of square, then he turned head toward the radar direction (see Fig. 38.7) and turned around by 90° (see Fig. 38.8). Figure 38.7 defines as 0°, Fig. 38.8 defines as 90°. Subject state is face up and face down. We can define face up as Up and face down as Down. We compare each state (see Figs. 38.9, 38.10). Figures 38.9 and 38.10s x-axis mean time index, and y-axis mean voltage (V) of instantaneous amplitude of I-Q signal. These figures show 90° is able to detect aspiration than 0°.

Fig. 38.7 Experimental condition -0°-

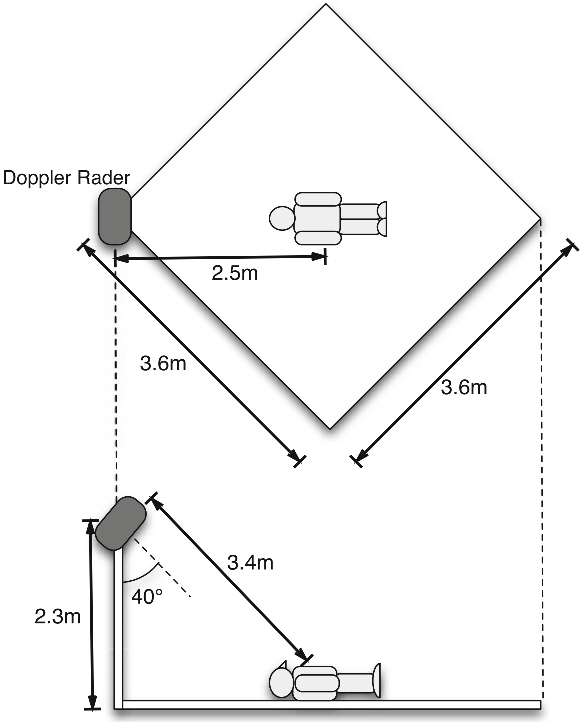


Fig. 38.8 Experimental condition -90°-

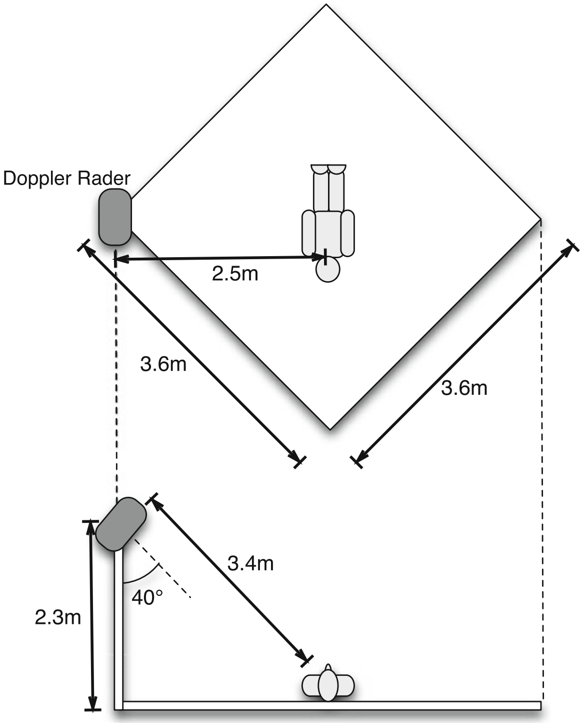
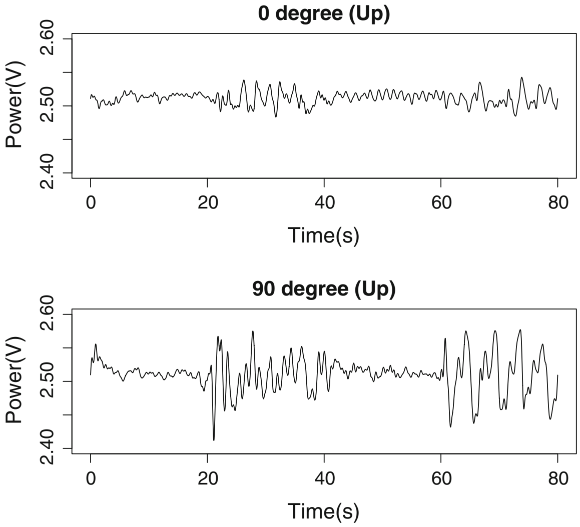


Fig. 38.9 Comparision 0° and 90° *-Up-*



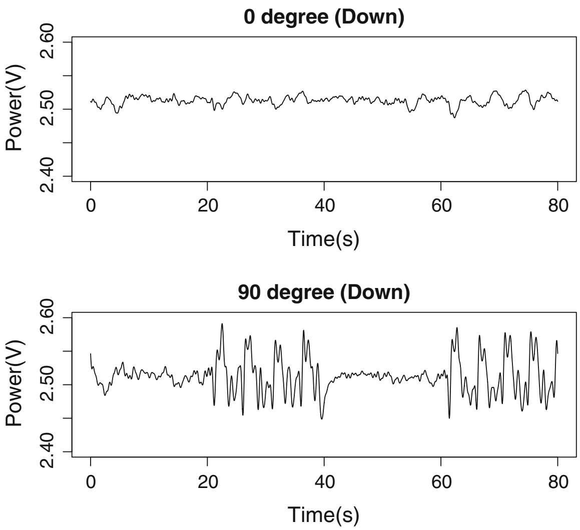


Fig. 38.10 Comparision 0° and 90° *-Down-*

We set the number of subjects was 4, the radar frequency was 1,000 Hz, and we measured 80 s. As measurement, subject held his aspiration for 20 s, then He took in a aspiration for 20 s. We made twice the flow of the above as a dataset.

38.6 Result

We discuss the results of our model. We apply SVM classifier to the estimated parameters to perform aspiration/apnea detection. We used the data which cut out 15 s of data each 20 s of apnea and aspiration data. To evaluate our model, we apply to leave-one-out cross validation test.

Tables 38.2, 38.3, 38.4 and 38.5 show average recognition rates of each experiments (Table 38.1). Tables 38.2 and 38.3 mean recognition rates of “face up” and “face down” at 0° respectively. Likewise, Table 38.4 and 38.5 mean recognition rates of “face up” and “face down” at 90° respectively. In Tables 38.2–38.5, “Hold” and “Take” means “Hold subject’s aspiration” and “Take aspiration” respectively.

Table 38.1 Summary of Tables 38.2–38.5

	State	Direction (°)
Table 38.2	Up	0
Table 38.3	Down	0
Table 38.4	Up	90
Table 38.5	Down	90

Table 38.2 Result of recognition rate (%) -1-

	Hold	Take
Hold	85.0	16.7
Take	15.0	83.3

Table 38.3 Result of recognition rate (%) -2-

	Hold	Take
Hold	76.5	40.0
Take	23.5	60.0

Table 38.4 Result of recognition rate (%) -3-

	Hold	Take
Hold	93.0	15.0
Take	7.0	85.0

Table 38.5 Result of recognition rate (%) -4-

	Hold	Take
Hold	93.8	9.5
Take	6.2	90.5

38.7 Conclusion and Future Work

According to experimental results, 90° case is easy to detect aspiration than 0° case, because the irradiate area of subject's aspiration motion is large. We have the problem that it is apt to affect by the state or place of subject. We need to cope with the robust to the state or place of subject.

This study gets the only data which observed center of square to assume the eight-mat. Healthcare application needs high and robust recognition rate of subject's state. We need to investigate the various environments.

Acknowledgments This research was supported in part by Adaptable and Seamless Technology Transfer Program through Target-driven R&D, Japan Science and Technology Agency.

References

1. Aoki S, Onishi M, Kojima M, Fukunaga F (2005) Detection of a Solitude Senior's irregular states based on learning and recognizing of behavioral patterns (in Japanese). IEEJ Trans Sens Micromach 125:259–265

2. Inui S, Okusa K, Maeno K, Kamakura T (2012) Recognizing aspiration presence using model parameter classification from microwave Doppler signals. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering and computer science (2012) WCECS 2012, 24–26 October, 2012. USA, San Francisco, pp 509–512
3. Kubo H, Mori T, Sato T (2010) Detection of human motion and respiration with microwave Doppler sensor. Jpn Soc Med Biol Eng 48:595–603
4. Kubo H, Mori T, Sato T (2011) Respiration measurement with microwave Doppler sensor (in Japanese). 16th Robotics Symposia, pp 111–118
5. Kurata T (2013) Introduction to support vector machine (in Japanese) (2002) http://home.hiroshima-u.ac.jp/tkurita/lecture/svm.pdf. Cited 7 Jan 2013
6. Lien PH, Lin FL, Chuang HR (2009) Computer simulation of the RF system effects on a millimeter-wave Doppler radar for human vital-signal estimation. In: Proceedings of the 6th European radar conference, pp 465-468
7. Naoi T, Maeda N, Iwama S (2005) Non-contact cardiac-beat detection using time-difference of microwave Doppler signal (in Japanese). IEICE Technical Report MBE 105, pp 5–8
8. Petrochilos N, Rezk M, Host-Madsen A, Lubecke V, Boric-Lubecke O (2007) Blind separation of human heartbeat and breathing by the use of a Doppler radar remote sensing. ICASSP, pp 333–336
9. Sekine M, Maeno K, Nozaki M (2009) Activity and state recognition without body-attached sensor using microwave Doppler sensor (in Japanese). IPSJ SIG Technical Report, 24th UBI, 10, pp 1-8
10. Tanigawa S, Nishihara H, Kaneda S (2008) The detection of the chewing with microwave Doppler sensor (in Japanese). 70th information processing society of Japan, 4, pp 273–274
11. Zhou Q, Liu J, Host-Madsen A, Boric-Lubecke O, Lubecke V (2006) Detection of multiple heartbeat using Doppler rader. ICASSP, pp 1160–1163

Chapter 39
Classification of Hyperspectral Images Using Machine Learning Methods

Bolanle Tolulope Abe, Oludayo O. Olugbara and Tshilidzi Marwala

Abstract Mixed pixels problem has significant effects on the application of remote sensing images. Spectral unmixing analysis has been extensively used to solve mixed pixels in hyperspectral images. This is based on the knowledge of a set of unidentified endmembers. This study used pixel purity index to extract endmembers from hyperspectral dataset of Washington DC mall. Generalized reduced gradient (GRG) a mathematical optimization method is used to estimate fractional abundances (FA) in the dataset. WEKA data mining tool is chosen to develop ensemble and non-ensemble classifiers using the set of the FA. Random forest (RF) and bagging represent ensemble methods while neural networks and C4.5 represent non-ensemble models for land cover classification (LCC). Experimental comparison between the classifiers shows that RF outperforms all other classifiers. The study resolves the problem associated with LCC by using GRG algorithm with supervised classifiers to improve overall classification accuracy. The accuracy comparison of the learners is important for decision makers in order to consider tradeoffs in accuracy and complexity of methods.

B. T. Abe (✉)
School of Electrical and Information Engineering, University of the Witwatersrand,
Johannesburg, South Africa
e-mail: abe_tolulope@yahoo.com

B. T. Abe
Department of Electrical Engineering, Tshwane University of Technology, Pretoria,
South Africa

O. O. Olugbara
Depatment of Information Technology, Durban University of Technology, Durban,
South Africa
e-mail: oludayoo@dut.ac.za

T. Marwala
University of Johannesburg, Johannesburg, South Africa
e-mail: tmarwala@uj.ac.za

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_39,

Keywords Accuracy · Classifier · Ensemble · Hyperspectral · Image · Optimization

39.1 Introduction

Recent developments in sensor technology have resulted to development of hyperspectral instruments. Each pixel in hyperspectral data cube is linked to spectral signature that uniquely describes the materials within the pixels. Such recognition provides great advantage for detecting minerals, urban planning and vegetation studies, monitoring and management of environment among others [1]. Hence, accurate classification of remote sensing images is crucial to achieve these advantages [2]. The challenge is that image objects are usually embedded in a single pixel and cannot be detected spatially [3, 4]. The traditional spatial based image processing techniques cannot be used.

In remote sensing technology, diversities of earth objects are present in the direct view of sensors because of the complexity of target objects and the limited spatial resolution of remote sensors [4]. The information present in a particular pixel of a remote sensing image is a mixture of information on various ground objects, resulting into mixed pixels [4–7]. The existence of mixed pixels has major effects on some practical applications of remote sensing images such as image classification, object detection and information extraction. It is an important task in remote sensing study to discover objects and corresponding quantity present in the mixed pixel so as to proffer solutions to the mixed pixel problem in remotely sensed imagery.

This work investigates the challenge of land cover classification of hyperspectral images by exploring a linear spectral mixture analysis (LSMA) approach. The objectives of the work are to (1) extract endmembers which is a collection of spectrally pure constituent spectral [4, 8] using Washington DC mall imagery dataset [9], (2) explore generalized reduced gradient (GRG) optimization algorithm to estimate the fractional abundance in the dataset thereby obtaining numeric values for land cover classification [4, 10, 11], (3) experimentally compare the performance of non-ensemble against ensemble supervised machine learning classification techniques. The non-ensemble classifiers used are neural networks and C4.5, while bagging and random forests represent the ensemble classifiers. The classifiers are used to examine the suitability of GRG algorithm for solving land cover classification problem. Supervised machine learning technique is based on task of inferring a function from labeled training data. The learning scheme is able to automatically extract information using computational and statistical approach for prediction because of its ability to improve classification performance [12, 13].

Machine learning based ensemble methods are essentially a multi-classifier system, implying that their functionality is dependent on a collection of classifiers—an ensemble of classifiers to get an accurate result Research has shown that ensemble generates better classification accuracy results than the individual classifier making up the ensemble [4, 14, 15].

This paper is succinctly organized as follows. In Sect. 39.2, the problem statement is described. In Sect. 39.3, machine learning methods used for experiments are briefly described. In Sect. 39.4, the design method of the work is discussed. In Sect. 39.5, the results of endmember determination classification accuracy of the learning methods are presented. In Sect. 39.6, the main conclusion achieved is summarized.

39.2 Problem Statement

The problem of land cover classification can be formulated as a linear spectral unmixing (LSU) problem. The LSU is a sub-pixel classification process that decomposes mixed pixels and determines the combination of fractional abundances. Based upon the linear unmixing model assumptions, each pixel at spatial coordinate (i,j) for a particular band in a remotely sensed hyperspectral image (I) with N number of bands can be formally expressed as [4, 7, 9]:

$$y(i,j) = \sum_{p=1}^{z} a_p(i,j).e_p + n(i,j) \tag{39.1}$$

The component $y(i,j)$ is a measured reflectance value at the spectral coordinate (i, j), e_p is the spectral response of the pth endmember, $a_p(i,j)$ is the fractional abundance of the pth endmember, $n(i,j)$ is the spectral band error and z is the total number of endmembers. For linear LSU, each image pixel is a mixture of various endmembers and spectrum recorded by sensor is a linear combination of endmembers spectral [23].

Equation (39.1) operates under two physical constraints on fractional abundances to account for the full composition of a mixed pixel [4, 7]. These are:

1. Non-negativity constraint, all abundances should be positive, that is

$$a(i,j)_p \geq 0, \ \forall p = 1, 2, \ldots, z \tag{39.2}$$

2. The abundances sum to one constraint,

$$\sum_{p=1}^{z} a_p(i,j) = 1 \tag{39.3}$$

The endmembers e_p, where p = 1, 2, …, z can be extracted from the image (I) by using a certain algorithm such as pixel purity index (PPI) and automated morphological endmember extraction (AMEE) [4, 7, 10] before the equation can be solved for a set of fractional abundances. Extant works on LSU problem

[4, 7–9] have explored the least square method to estimate a set of fractional abundances as follows.

$$a_{LSU}(i,j) = (e^T e)^{-1} e^T y(i,j) \tag{39.4}$$

Equation (39.4) can only satisfy the sum to unity of abundances, but the non-negativity of fractional abundances cannot always be guaranteed. The results obtained by the least square method therefore, are generally not optimal in terms of material quantification [4, 7].

In order to find a set of fractional abundances that simultaneously satisfy these two constraints, the following fully constraint linear spectral unmixing GRG optimization problem formulation has to be solved to minimize the spectral band error $n(i,j)$.

Minimize:

$$\begin{aligned} &a(i,j) \in \\ &\Delta\{(y(i,j) - a(i,j).e)^T (y(i,j) - a(i,j).e)\} \end{aligned} \tag{39.5}$$

Subject to:

$$\Delta_1 = \left\{ a(i,j) | \sum_{p=1}^{z} a_p(i,j) - 1 = 0 \right\} \tag{39.6}$$

$$\Delta_2 = \{a(i,j) | 0 \le a_p(i,j) \le 1\} \tag{39.7}$$

In Eq. (39.5), the expression $(y)^T$ represents the transpose vector of the vector (y). In order to solve Eqs. (39.5–39.7) by applying GRG algorithm [4, 12, 13], to obtain an estimate of a set of fractional abundances, the PPI algorithm is first applied to extract endmembers from the hyperspectral image. The PPI method efficiently handles hyperspectral imagery as it provides a convenient and physically motivated decomposition of an image in terms of a relatively few components [4, 24].

After a set of endmembers $e = \{e_p\}_{p=1}^{z}$ is determined, the corresponding fractional abundances $a(i,j) = \{a_p(i,j)\}_{p=1}^{z}$ in a certain pixel vector $y(i,j)$ on the image I is estimated by using the GRG algorithm.

39.3 Learning Methods

Neural network methods are general classifiers that can handle problems with lots of parameters and can classify objects even when the distribution of object in n-dimensional parameters space is very complex. Research activities have established that neural networks are capable alternative to numerous conventional methods [4, 16–18]. They are data driven self-adaptive techniques that adjust themselves to data under investigation without any explicit specification of functional or distributional form. They are also universal functional approximators that can approximate any function with arbitrary accuracy [4]. In addition, neural networks are nonlinear models that make them flexible in modeling real world complex relationships. Various neural networks are available for classification purposes [4, 18], but this work focuses on multilayer perceptron (MLP) that uses back propagation to classify instances. The nodes in the network are all sigmoid.

C4.5 algorithm creates a decision tree classifier to predict membership cases of categorical dependent variable from measures on one or more variable. It uses information gain ration matrix for classification. C4.5 uses significance of statistic of error to trim branch and uses probability weighting to deal with feature loss during training period [19]. The algorithm works in a way that each node of the decision tree matches an attribute and individual arc matches a value range of the attribute. The value of the expected attribute is known by the path from the root to individual leaf. The highest attribute is allocated to each node. This aims at associating the attribute to reduce the data entropy to a node [20].

Bagging also known as bootstrap aggregating is a leading widely used technique of selecting sets for ensemble classifiers. The invention of bagging has its root from attempts to reduce the error variance to assists classification accuracy. The algorithm operates by creating new training sets using resampling methods from the original dataset n (the number of samples in the original training data) times, randomly with replacement. The sample been chosen will not be removed from the dataset in the next draw. Some of the training samples will be selected more than once while some samples will not be chosen at all in a new set. The classifiers are ensemble using majority vote and vote of each classifier carries the same weight [21]. For this experiment, the ensemble classifier is made up of ten decision table majority base classifiers.

Random Forest is a supervised ensemble classifier that builds several decision trees randomly [4, 22] for classification of multisource remote sensing, geographic data and hyperspectral imaging. RF searches only a random subset of the features for a split at each node to minimize the correlation between the classifiers in the ensemble. It selects a set of features randomly and creates an algorithm with a bootstrapped sample of the training data [4, 22]. The technique offers a potential benefit in that it is insensitive to noise or overtraining because resampling is independent of the weighting scheme employed. For our experiment, 10 trees were constructed. Out of bag error was 0.0605 while considering 192 random features.

39.4 Method

The method of design of this work entails arrangement of steps that the input hyperspectral image undergoes for its land covers to be classified into one of the desired multiple classes. The input data has to be taken through four steps to obtain the desired classification result. Figure 39.1 shows the block diagram of the land cover classification process implemented in this study.

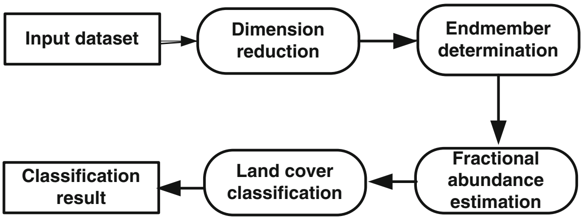


Fig. 39.1 Hyperspectral image classification procedure

Before discussing the vital steps used in our design method, the input dataset is first introduced.

39.4.1 Dataset

Figure 39.2 shows the input airborne hyperspectral image of Washington D.C. dataset [9]. The sensor used measured pixel response in 210 bands in the 0.4–2.4 µm region of the visible and infrared spectrum. It has 1,208 scan lines with 307 pixels in each scan line, which is approximately 150 MB. Bands in the 0.9 and 1.4 µm region where the atmosphere is opaque have been omitted. The remaining 191 spectral bands are used for this study. The dataset shown in Fig. 39.2 contains seven ground cover types namely Roofs, Street, Path, Grass, Trees, Water and Shadow.

Accompany the dataset is a copy of the file labeled *dctest.project,* which describes the land cover types (also refer to as class labels) used for the experimental procedure.

39.4.2 Image Dimension Reduction

Dimension reduction as applied to hyperspectral image aims at reducing the number of spectral bands in the image. This is done to map the data into lower dimension from higher dimension at the same time preserve the main features of the original data. The process is carried out to reduce the time used during the processing of the hyperspectral data. The algorithm is designed to reduce error by

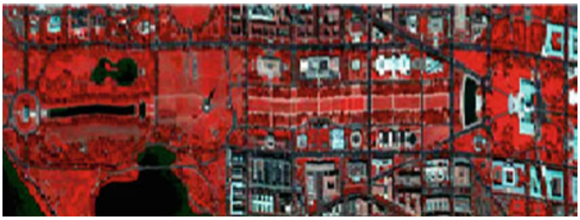


Fig. 39.2 Hyperspectral image of Washington D. C. mall [9]

finding minimum representation of the original image that adequately keeps the original information for successful unmixing in the lower dimension [25]. Among various algorithms normally used for dimension reduction are principal component analysis (PCA) and maximum noise fraction (MNF) transform. The aim is to ease computational complexity and for compact information in transformed components. MNF is used in this study because it is more effective than PCA [4, 26].

39.4.3 Endmember Determination

An endmember is known as a spectrally pure pixel that portrays various mixed pixel in an image [4, 27]. The method of feature selection involves identifying the most discriminative measurements out of a set of H potentially useful measurements, where h $\leq$ H. Endmember extraction has been widely used in hyperspectral image analysis to significantly improve spatial and spectral resolution provided by hyperspectral imaging sensor also known as imaging spectrometry [26]. Identification of image endmember is a crucial task in hyperspectral data exploitation, especially classification [4, 8]. After endmembers selection, various methods can be used to construct their special distribution, associations and fractional abundances. For real hyperspectral data, various algorithms have been developed to execute the task of locating appropriate endmembers. These include pixel purity index (PPI), N-FINDR and automatic morphological endmember extraction (AMEE) [4, 26].

This work applies the PPI algorithm [4, 28], which is available in the environment for visualizing images (ENVI) to determine endmembers from the hyperspectral image. The choice of the algorithm is motivated by its publicity in ITTVIS (http://www.ittvis.com/) ENVI software that was originally developed by analytical imaging and geophysics (AIG) [29]. PPI generates a large number of n-dimensional vectors called “skewer” [28] through the dataset. N-FINDR fully automated method locates the set of pixels with the largest possible volume by “inflating” a simplex within the image data. “Noise whitening” and dimensionality reduction are performed using MNF transform to generate the endmembers [27]. Then pixel purity score is obtained in the image cube by producing lines in the n-dimension space containing the MNF-transformed data. The spectral points are projected on the lines and the points at the extremes of each line are counted. Bright pixels in the PPI image generally are image endmembers. The highest-valued of these pixels are input into the n-dimensional visualizer for the clustering process that develops individual endmember spectral.

39.4.4 Fractional Abundance Estimation

After determining the endmembers using PPI procedure, per pixel fractional abundances of various materials is estimated using GRG optimization method. This study presents six endmember models to characterize the land cover structure which are Roofs, Street, Path, Grass, Trees, Water and Shadow. Normalized numerical values of the fractional abundant generated were calculated from the spectral signatures of the land cover label signatures. The values obtained were used to train the random forests and neural networks classifiers for land cover classification [4].

39.4.5 Land Cover Classification

Neural networks (NN), C4.5, Bagging and Random forests (RF) machine learning classification methods are experimentally compared to examine their performances in the field of land cover classification. The Waikato environment for knowledge analysis (WEKA) [30] data mining software is selected as a tool to build the classifiers from a training dataset of 3,355 instances and 191 band features.

39.5 Results

This section presents the results and discussion of our experiment for fractional endmember determination and classification accuracy of the classifiers investigated.

39.5.1 Fractional Endmember

The first experiment performed aimed to obtain endmembers from image dataset using the ENVI software application. The MNF transformation of the hyperspectral image was performed for dimension reduction. The next stage is to select a set of endmembers by applying the PPI algorithm on the extracted region of interest (ROI) pixels. Figure 39.3 shows this result, wherein the extreme pixels corresponding to the endmembers in each projection are recorded and total number of times each pixel is marked as extreme is noted. A threshold value of 1 is used to define how many pixels are marked as extreme at the ends of the projected vector.

Table 39.1 displays the land cover classes and the number of pixels extracted from the original image based on the ROI. The values of these pixels are input into the ENVI visualizer for the clustering process that develops individual endmember

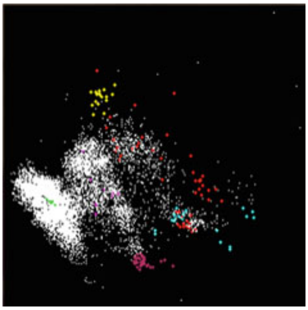


Fig. 39.3 Purest pixels occur at edges of the projected vector

Table 39.1 Number of pixels extracted from ROI

Classes	Number of pixels
Roof	724
Paths	211
Water	703
Street	404
Trees	398
Shadow	97
Grass	818

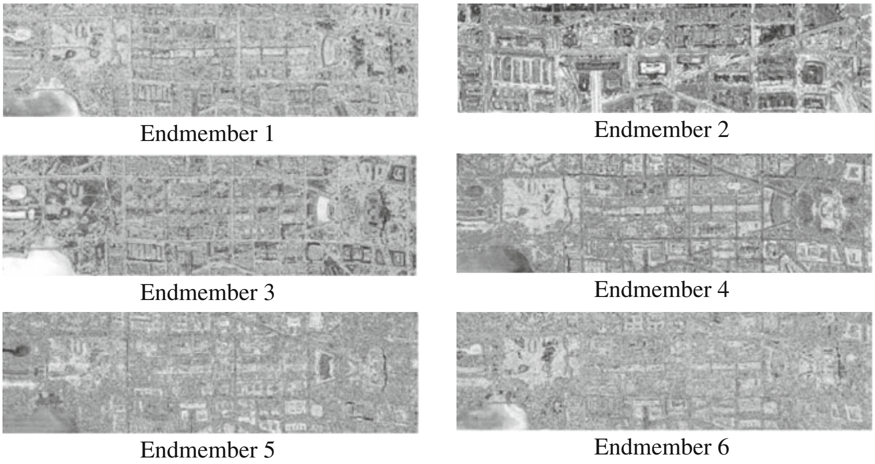
spectral. The pixels extraction mechanism enables the image spectral to accurately account for any errors in atmospheric correction.

The estimated number of spectral endmembers and their corresponding spectral signatures are obtained using ENVI visualizer. Figure 39.4 shows six fractional endmembers of the image generated from the PPI method.

At the completion of specified iterations, a PPI image is created in which the value of each pixel corresponds to the number of times that a pixel was recorded as extreme. The bright pixels in the PPI image are generally the image endmembers to characterize the land cover structure.

39.5.2 Classification Accuracy

NN, C4.5, Bagging and RF learning classifiers are evaluated using the method of error confusion matrix, which is also known as contingency table, a representation of the entire classification results. According to [31], error confusion matrix can be used to compute overall accuracy and individual class label accuracy. The error confusion matrix is a widely accepted method to report error of raster data and to assess classification accuracy of a classifier. The matrix expresses the number of sample units allocated to each land cover type as compared to the reference data [that is it shows the number of false positives (FP) and false negatives (FN)].



Endmember 1

Endmember 2

Endmember 3

Endmember 4

Endmember 5

Endmember 6

Fig. 39.4 Fraction images for each endmember

Table 39.2 Neural networks error confusion matrix

a	b	C	d	e	f	g	classified as
724	0	0	0	0	0	0	a = Roofs
1	210	0	0	0	0	0	b = Paths
1	0	699	0	0	3	0	c = Water
1	0	2	401	0	0	0	d = Streets
0	0	0	0	398	0	0	e = Trees
0	0	17	0	0	80	0	f = Shadow
0	0	0	0	0	0	818	g = Grass

The diagonal of the matrix designates agreement between reference data and interpreted land cover types [4, 32].

Table 39.2 records result of error confusion matrix for performance of NN. From the table, it can be observed that roofs, trees and grass are 100 % classified while other land cover classes have some of their pixels misclassified.

Table 39.3 presents the result of error confusion matrix for performance of C4.5 classifier. From the table, it can be deduced again that roofs and grass are 100 % classified while other land cover classes have some of their pixels misclassified.

As for bagging technique, the result obtained from the method as seen in Table 39.4 reveals that none of the class labels is 100 % classified.

Table 39.5 shows the result of error confusion matrix for performance of RF classifier. This result shows that roofs, paths, water, streets, trees and grass have 100 % classification accuracy because none of their pixel's member is misclassified while shadow has only one of the pixels' members misclassified. The result reveals that the performance of RF classifier considering per class classification accuracy is outstanding.

Table 39.3 C4.5 error confusion matrix

a	b	C	d	e	f	g	<– classified as
724	0	0	0	0	0	0	ǀ a = Roofs
2	209	0	0	0	0	0	ǀ b = Paths
0	0	686	1	0	16	0	ǀ c = Water
3	0	0	401	0	0	0	ǀ d = Streets
0	0	0	0	396	0	2	ǀ e = Trees
0	0	0	0	0	96	1	ǀ f = Shadow
0	0	0	0	0	0	818	ǀ g = Grass

Table 39.4 Bagging error confusion matrix

a	b	C	d	e	f	g	<– classified as
710	1	4	9	0	0	0	ǀ a = Roofs
1	207	0	0	0	0	3	ǀ b = Paths
0	0	691	0	0	12	0	ǀ c = Water
10	1	0	389	0	4	0	ǀ d = Streets
0	0	0	0	397	0	1	ǀ e = Trees
0	0	11	4	0	82	0	ǀ f = Shadow
0	0	0	0	1	0	817	ǀ g = Grass

Table 39.5 Random forests error confusion matrix

a	b	C	d	e	f	g	classified as
724	0	0	0	0	0	0	a = Roofs
0	211	0	0	0	0	0	b = Paths
0	0	703	0	0	0	0	c = Water
0	0	0	404	0	0	0	d = Streets
0	0	0	0	398	0	0	e = Trees
0	0	1	0	0	96	0	f = Shadow
0	0	0	0	0	0	818	g = Grass

Generally, all classifiers performed excellently well. Considering individual class label, RF produces the highest level of classification accuracy per class label as compared to the remaining classifiers. The entire accuracy assessment procedure is that the error confusion matrix must be a representative of the entire area mapped from the remotely sensed data [4, 32]. The overall accuracy for correctly classified instances, incorrectly classified instances, unclassified instances and Kappa statistic are identified from the error confusion matrices [4, 32].

If all non-major diagonal elements of error confusion matrix are zero, then it means no area in the map has been misclassified or correctly classified instances (CCI) and the map accuracy is 100 %. Otherwise there are certain percentages of incorrectly classified instances (ICI) [4]. In our experiment, RF as compared to other classifiers used has only 1 instance misclassified, while NN, Bagging has 25, 25, 62 instances misclassified respectively.

Table 39.6 Performance result summary

C	CCI	ICI	KS	MAE	RMSE	RAE (%)	RRSE (%)	Accuracy (%)
RF	3354	1	0.9996	0.0015	0.0176	0.6568	5.1615	99.9702
NN	3330	25	0.9909	0.003	0.0379	1.2835	11.1095	99.2548
C4.5	3330	25	0.9909	0.0031	0.0393	1.3265	11.5178	99.2548
BD	3293	62	0.9774	0.0455	0.1	19.4796	29.2879	98.152

The following equations are used for calculating the performance measure of the classification procedure [33]:

$$\text{Mean absolute error (MAE)} \quad \frac{|q_1 - b_1| + \cdots + |q_n - b_n|}{n}$$

$$\text{Root mean square error (RMSE)} \quad \sqrt{\frac{(q_1 - b_1)^2 + \cdots + (q_n - b_n)^2}{n}}$$

$$\text{Relative absolute error (RAE)} \quad \frac{|q_1 - b_1| + \cdots + |q_n - b_n|}{\left|b_1 - \bar{b}\right| + \cdots + \left|b_n - \bar{b}\right|}$$

$$\text{Root relative squared error (RRSE)} \quad \sqrt{\frac{(q_1 - b_1)^2 + \cdots + (q_n - b_n)^2}{(q_1 - \bar{b})^2 + \cdots + (q_n - \bar{b})^2}}$$

where q represents predicted values and b is the real values. Table 39.6 shows the result of performance measure of the classifiers and overall accuracy classification.

Kappa statics (KS) coefficient of agreement is a measure of how well the accuracy of the classifier compares with the reference or ground truth data [34]. It ranges from 0 to 1, with 0 implying no agreement between the classified land cover and ground truth and 1 indicates complete agreement. If the Kappa statics varies between 0.40 and 0.59, it is assumed as moderate, 0.60–0.79 is considered as substantial and above 0.80 is considered as outstanding [35]. According to these results, there are no unclassified instances during classification procedures and overall classification accuracies of learning classifiers are seen to be comparable.

Validation of the results: to validate the performance of the classifiers, our test set was run through the models. Table 39.7 illustrates the outcome of the validation test as compared with the classification accuracy results obtained from each model.

The results show that accuracy results obtained from the models are pretty close, which indicate that the classifiers will not breakdown with unknown data or when future data are applied [13, 15, 34, 36]. It can be deduced from the predictions that RF outperformed all other classifiers. In addition, RF is more computational effective as compared to others.

Table 39.7 Comparison of accuracy and validation results

Classifier	Classification accuracy (%)	Test validation (%)
Neural network	99.2548	99.3722
C4.5	99.2548	98.8341
Bagging	98.152	97.6682
Random forest	99.9702	97.8475

39.6 Conclusion and Future Work

This study aimed to establish performance comparison between ensemble and non-ensemble classifiers for land cover classification. The performance assessment was done, giving overall accuracy and error confusion matrix. Experimental results demonstrate that neural network, C4.5, bagging and random forest based land cover classification systems significantly improve overall accuracy. The validation results also confirm that the models will not breakdown when future data are applied to the learning methods. As a result, the classifiers can significantly contribute to land cover classification system as a source of analysis and increase its accuracy. The comparability and high accuracy performance of the ensemble and non-ensemble systems indicates that GRG method introduced in this study is effective for solving a LSU problem of land cover classification.

Future work is to implement generalize reduce gradient (GRG) optimization approach on different datasets to authenticate the viability.

Acknowledgments This work was supported by University of Johannesburg, South Africa, Tshwane University of Technology, South Africa and University of the Witwatersrand, Johannesburg. South Africa.

References

1. Xie Y, Sha Z, Yu M (2008) Remote sensing imagery in vegetation mapping: a review. J Plant Ecol 1:9–23
2. Shaw GA, Burke HK (2003) Spectral imaging for remote sensing. Lincoln Lab J 14(1):3–28
3. Chang CI, Heinz DC (2000) Constrained subpixel target detection for remotely sensed imagery. IEEE Trans Geosci Remote Sens 38(3):1144–1159
4. Abe BT, Olugbara OO, Marwala T (2012) Hyperspectral image classification using random forest and neural network. In: Proceedings of the world congress on engineering and computer science 2012, Lecture notes in engineering and computer science, 24–26 October, San Francisco, USA, pp 522–527
5. Sanchez S, Martin G, Plaza A, Chang C (2010) GPU implementation of fully constrained linear spectral unmixing for remotely sensed hyperspectral data exploitation. In: Proceedings SPIE satellite data compression, communications, and processing VI, 2010, vol 7810, pp 78100G-1–78100G-11
6. Iordache M-D, Bioucas-Dias JM, Plaza A (2011) Sparse unmixing of hyperspectral data. Geosci Remote Sens IEEE Trans 49(6):2014–2039
7. Zhang B, Sun X, Gao L, Yang L (2011) Endmember extraction of hyperspectral remote Sensing images based on the ant colony optimization (ACO) algorithm. Geosci Remote Sens IEEE Trans 49(7):2635–2646
8. Martinez PJ, Perez RM, Plaza A, Aguilar PL, Cantero MC, Plaza J (2006) Endmember extraction algorithms for hyperspectral images. Ann Geophy 49(1):93–101
9. Landgrebe DA (2003) Signal theory methods in multispectral remote sensing. Wiley, Hoboken
10. Abadie J, Carpentier J (1969) Generalization of the Wolfe reduced gradient method in the case of non-linear constraints. In: Fletcher R (ed) Optimization. Academic Press, London, pp 37–47

11. Lasdon LS, Fox RL, Ratner MW (1974) Nonlinear optimization using the generalized reduced gradient method. Revue française d' automatique, d' informatique et de recherché, 1974, issue 3, pp 73–103
12. Maree R, Stevens B, Geurts P, Guern Y, Mack P (2009) A machine learning approach for material detection in hyperspectral images. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition workshops, 2009, pp 106–111
13. Okori W, Obua J (2011) Machine learning classification technique for famine prediction. In: Proceedings of the world congress on engineering, 2011, vol II, July 6–8 2011, London, U.K, pp 991–996
14. Abe B, Gidudu A, Marwala T (2010) Investigating the effects of ensemble classification on remotely sensed data for land cover mapping. In: Proceedings of IEEE international conference on geoscience and remote sensing symposium (IGARSS), 2010, pp 2832–2835
15. Govindarajan M, Chandrasekaran RM (2012) Intrusion detection using ensemble of classification methods. In: Proceedings of the world congress on engineering and computer science 2012, Lecture notes in engineering and computer science, 24–26 October 2012, San Francisco, USA, pp 459–464
16. Benediktsson JA, Swain PH, Ersoy OK (1990) Neural network approaches versus statistical methods in classification of multisource remote sensing data. IEEE Trans Geosci Remote Sens 28:540–552
17. Ramírez-Quintana JA, Chacon-Murguia MI, Chacon-Hinojos JF (2012) Artificial neural image processing applications: a survey. Eng Lett 20:1–68 ([Online] Available http://www.engineeringletters.com/issues_v20/issue_1/EL_20_1_09.pdf)
18. Mekanik F, Imteaz MA (2012) A multivariate artificial neural network approach for rainfall forecasting: case study of Victoria, Australia. In: Proceedings of the world congress on engineering and computer science 2012, Lecture notes in engineering and computer science, 24–26 October 2012, San Francisco, USA, pp 557–561
19. Wang P, Zhang J, Jia W, Lin Z (2008) A study on decision tree classification method of land use/land cover -taking tree counties in Hebei province as an example. In: Proceedings of earth observation and remote sensing applications, 2008, international workshop on 2008, pp 1–5
20. Pinho CMD, Silva FC, Fonseca LMC, Monteiro AMV (2008) Intra-urban land cover classification from high-resolution images using the C4.5 algorithm. In: Proceedings of the international archives of the photogrammetry, remote sensing and spatial information sciences, 2008, vol 37(B7), pp 695–699
21. Breiman L (1996) Bagging predictors. Mach Learn 24(2):123–140
22. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
23. Kärdi T (2007) Remote sensing of sensing of urban areas: linear spectral unmixing of landsat thematic Mapping images acquired over Tartu (Estonia). In: Proceedings of the Estonian academy of sciences: biology, ecology, 2007, vol 56, no 1, pp 19–32
24. Theiler J, Lavenier D, Harvey N, Perkins S, Szymanski J (2000) Using blocks of skewers for faster computation of pixel purity index. In: Proceedings of the SPIE international conference on optical science and technology, 2000, no 4132, pp 61–71
25. Keshava N, Mustard JF (2002) Spectral unmixing. IEEE Signal Process Mag 19(1):44–57
26. Chaudhry F, Wu C, Liu W, Chang C-I, Plaza A (2006) Pixel purity index-based algorithms for endmember extraction from hyperspectral imagery. In: Proceedings of recent advances in hyperspectral signal and image processing, C.-I Chang, Ed. Trivandrum, India: Research Signpost, 2006, no 3, pp 31–61
27. Plaza A, Martinez P, Perez R, Plaza J (2004) A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. IEEE Trans Geosci Remote Sens 42(3):650–663
28. Gonzalez C, Resano J, Mozos D, Plaza A, Valencia D (2010) FPGA implementation of the pixel purity index algorithm for remotely sensed hyperspectral image analysis. EURASIP J Adv Signal Process 969806:1–13
29. Boardman JW, Biehl LL, Clark RN, Kruse FA, Mazer AS, Torson J (2006) Development and implementation of software systems for imaging spectroscopy. In: Proceedings of IEEE

international conference on geoscience and remote sensing symposium, 2006, July 31 2006–Aug 4 2006, pp 1969–1973

30. Garner SR (1995) WEKA: the Waikato environment for knowledge analysis. In: Proceedings of the NewZealand computer science research students conference, 1995, pp 57–64
31. Benediktsson JA, Swain PH, Ersoy OK (1990) Neural network approaches versus statistical methods in classification of multisource remote sensing data. IEEE Trans Geosci Remote Sens 28:540–552
32. Congalton RG (1991) A review of assessing the accuracy of classifications of remotely sensed data. Remote Sens Environ 37(1):35–46
33. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco, pp 189–199, 316–319, 337–423
34. Congalton R (1988) A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data. Photogram Eng Remote Sens 54(5):593–600
35. Kumar Y, Sahoo G (2012) Analysis of parametric and non parametric classifiers for classification technique using WEKA. Int J Inf Technol Comput Sci 7:43–49. doi:10.5815/ijitcs.2012.07.06 (Published Online July 2012 in MECS http://www.mecs-press.org/)
36. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I (2009) The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 2009, vol 11

Chapter 40
Evaluating the Effect of Single and Combined Climate Modes on Rainfall Predictability

Fatemeh Mekanik and Monzur Alam Imteaz

Abstract This study attempts to find the effect of past values of El Nino southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) on future rainfall. Victoria located at southeast Australia has been chosen as the case study. Many studies have tried to establish the relationships of these large-scale climate indices among the rainfalls of different parts of Australia; unlike the other regions no clear relationship can be found between each individual large-scale climate mode and Victorian rainfall. Past studies considering southeast Australian rainfall predictability could achieve a maximum of 30% predictability. This study looks into the lagged-time relationships of single and combined climate modes with Victorian spring rainfall using the nonlinear technique Artificial Neural Networks (ANN). Using these climate indices in an ANN model increased the model correlation up to 0.99, 0.98 and 0.30 in the testing set for the three case study stations of Horsham, Melbourne and Orbost in Victoria, Australia respectively. It seems that past values of IOD and ENSO both have a great effect on rainfall forecasting however the effect of IOD is higher in centre and west of Victoria compared to ENSO, while both ENSO and IOD seem to have a strong effect on the east side. This method can be used for other parts of the world where a relationship exists between rainfall and large scale climate modes which could not be established by linear methods.

Keywords ANN · ENSO · Forecast · IOD · Rainfall · Victoria

F. Mekanik (✉) · M. A. Imteaz
Center for Sustainable Infrastructure (CSI), Swinburne University of Technology,
Hawthorn, VIC 3122, Australia
e-mail: fmekanik@swin.edu.au

M. A. Imteaz
e-mail: mimteaz@swin.edu.au

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_40,

40.1 Introduction

Forecasting rainfall several months or seasons in advance can be beneficial for the management of water resources. Many studies have tried to find the relationships between large-scale climate modes and rainfall in different parts around the world using different linear and nonlinear methods [1–6].

It is believed that Australian rainfall is affected by several major climate patterns. The major drivers bringing rainfall over Australia which have been investigated by many researchers are, El Nino Southern Oscillation (ENSO), Indian Ocean Dipole (IOD) and Southern Annular Mode (SAM) [7]. ENSO is represented by two different type of indices; the Southern Oscillation Index (SOI) which is a measure of Sea Level Pressure (SLP) anomalies between Darwin and Tahiti and Sea surface temperature (SST) anomalies in equatorial Pacific Ocean noted as NINO3 (5°S–5°N, 150°–90°W), NINO3.4 (5°S–5°N, 170°–120°W) and NINO4 (5°S–5°N, 160°–150°W) [8]. The IOD is also a coupled ocean–atmosphere phenomenon in the equatorial Indian Ocean [9]. SAM is the major mode of atmospheric variability on the mid and high latitude of southern hemisphere. Even though SAM has been discovered to affect Australian seasonal rainfall [10] and daily rainfall [11], however SAM is not considered in this study due to its shorter range of available data. Many researchers have conducted different studies in different parts of Australia trying to establish the relationship between these climate modes and Australian rainfall [7, 8, 10–17]. Victoria located in southeast Australia is one of the regions that so far did not show good correlation of its rainfall and the climate modes. According to [18] in comparison to eastern Australia and particularly Queensland, past studies considering southeast Australian rainfall predictability could achieve a maximum of 30 % predictability. Other than the work of [19] which analyzed the combined impact of ENSO and SAM on Victorian rainfall, other studies focused only on finding the relationship between rainfall and a single driver. In the work of [17] the maximum correlation of 0.37 was achieved for spring rainfall with spring NINO4. On the other hand, the studies did not take into account the relationship between previous time lags of these drivers as a potential predictor for future rainfall. The majority of the mentioned studies used linear regression analysis or probability/categorical analysis (below median/above median) between rainfall and the simultaneous major climate modes. However, according to [20] a strong relationship between simultaneous climate modes and rainfall does not essentially mean that there is a lagged relationship as well. Kirono et al. [7] is one of the few accessible publications considering the relationship between Australian rainfall and 2 months average lag of different climate indices. Abbot and Marohasy [21] also used past values of climate indices, monthly historical rainfall data and atmospheric temperature for monthly and seasonal forecasting of rainfall in Queensland; however the climate indices they used were limited to SOI, DMI, PDO and NINO3.4 lagged by 1–2 months. Schepen et al. [20] used a Bayesian joint probability modeling approach for seasonal rainfall prediction; however their results were to some extent different from [7].

According to [19] Victorian rainfall variability is not driven by a single climate mode. Further to investigating the effect of different lagged-time climate indices on Victorian spring rainfall, this chapter also contributes to finding the lag relationship of combined climate indices and Victorian rainfall in order to have a better understanding of the variability of Victorian rainfall in regards to large scale climate modes. Thus, this study which is an extended and revised work of [22] is distinguished from previous studies by forecasting spring rainfall three consecutive years in advance by using the lag relationship of separate and combined climate indices. The combined ENSO-IOD sets include lagged NINO3-DMI, NINO4-DMI, NINO3.4-DMI and SOI-DMI, since there is no agreement on which of the ENSO indices can better represent this ocean-atmospheric phenomenon.

40.2 Data

40.2.1 Rainfall Data

Historical monthly rainfall data was obtained from the Australian Bureau of Meteorology for Horsham, Melbourne and Orbost, representing west, centre and east Victoria, Australia as a case study. Figure 40.1 shows the location details of the station considered in this study. Spring (September–November) rainfall was obtained from monthly rainfall data from January 1900 to December 2009 (www.bom.gov.au/climate/data/).

40.2.2 Climate Indices

In this study, monthly values of NINO3, NINO4, and NINO3.4 were used as indicators of ENSO. In addition to this SST related indices, Southern Oscillation Index (SOI) which is the SLP indicator of ENSO was also considered in this study. A measure of IOD is the Dipole Mode Index (DMI) which is the difference in average SST anomalies between the tropical Western Indian Ocean (10°S–10°N, 50°–70°E) and the tropical Eastern Indian ocean (10°S—Equator, 90°–110°E) [7]. ENSO and IOD indices were obtained from Climate Explorer website (http://climexp.knmi.nl/).

40.3 Artificial Neural Network

Many probabilistic and deterministic modeling approaches have been used by hydrologist and climatologist in order to capture rainfall characteristics. Conceptual and physically based models require an in depth knowledge of this complex

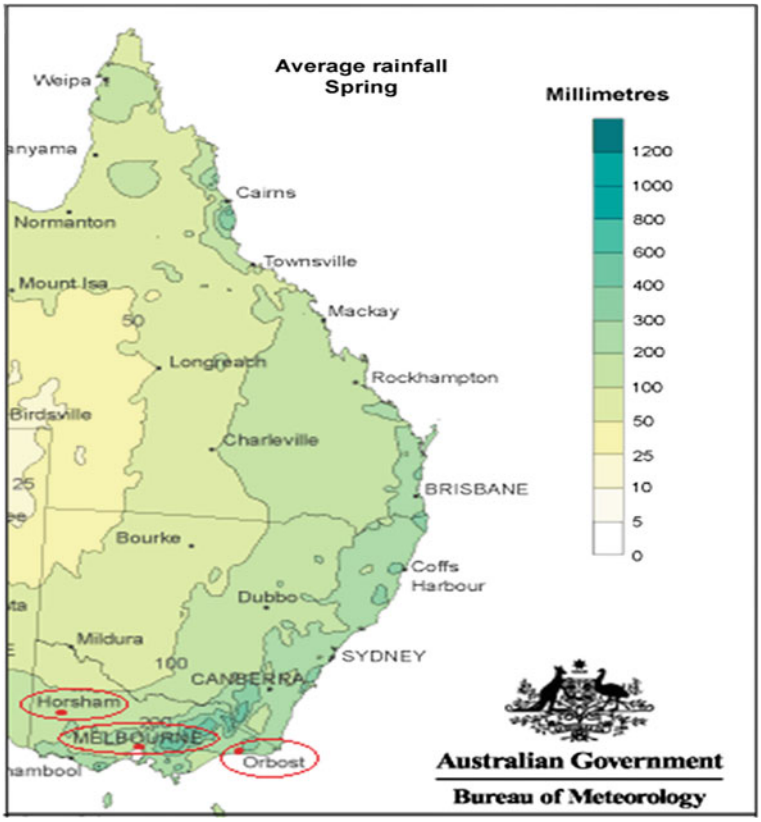


Fig. 40.1 Map of the study area (adopted from www.bom.gov.au/jsp/ncc/climate_averages/rainfall)

atmospheric phenomena; these models need a large amount of calibration data and they have to deal with over parameterization effect and parameter redundancy impact [23]. Artificial Neural Networks (ANN) is a mathematical model that has the ability to find the nonlinear relationship between input and output parameters without the need to solve complex partial differential equations [24]. The parameters for ANN modeling are basically network topology, neurons characteristics, training and learning rules. Multi-Layered Perceptrons (MLP) are feed-forward nets with one or more hidden layers between the input and output neurons (Fig. 40.2). The number of input and output neurons is based on the number of input and output variables. Basically, the input layer only serves as receiving the input data for further processing in the network. The hidden layers are a very important part in a MLP since they provide the nonlinearity between the input and output sets. More complex problems can be solved by increasing the number of hidden layers. The output neuron is the desired output of the model. The process of developing an ANN model is to find (a) suitable input data set, (b) determine the

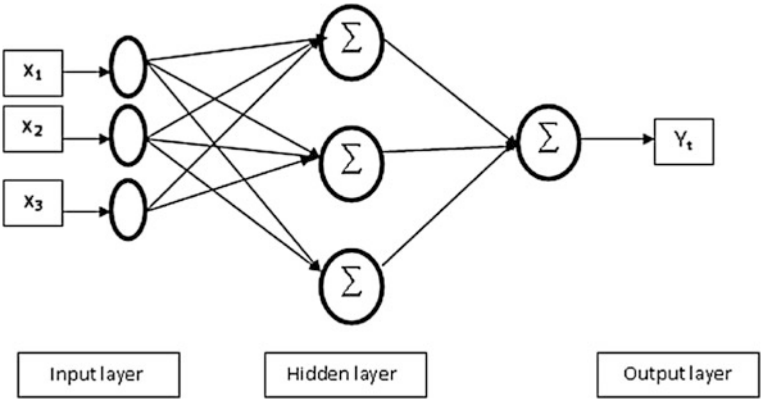


Fig. 40.2 A typical ANN architecture

number of hidden layers and neurons, and (c) training and testing the network. Mathematically, the network depicted in Fig. 40.2 can be expressed as follow:

$$Y_t = f_2\left[\sum_{j=1}^{J} w_j f_1\left(\sum_{i=1}^{I} w_i x_i\right)\right] \tag{40.1}$$

where Y_t is the output of the network, x_i is the input to the network, w_i and w_j are the weights between neurons of the input and hidden layer and between hidden layer and output respectively; f_1 and f_2 are the activation functions for the hidden layer and output layer respectively. According to [25] if extrapolating beyond the range of the training data is needed it is recommended to use sigmoiadal-type transfer functions in the hidden layers and linear transfer functions in the output layer. In this study f_1 is considered tansigmoid function which is a nonlinear function and f_2 is considered the linear purelin function defined as follow:

$$f_1 = \frac{2}{(1 + \exp(-2x))} - 1 \tag{40.2}$$

$$f_2(x) = x \tag{40.3}$$

The input set was chosen based on a priori knowledge of the predictors (e.g. the climate modes). ENSO cycle starts at April–May of a year and continues until March–April of the following year [16]. Spring rainfall occurs in the middle of the ENSO phase (positive, negative or neutral), thus the authors chose to take Dec$_n$-1-Aug$_n$ monthly values of ENSO indices (“n” being the year for which spring rainfall is being predicted) as inputs to find the maximum effect of ENSO on spring rainfall. This means that nine months lags of ENSO has been considered as potential predictors. On the other hand, according to [9] the IOD occurs in May to November and it matures in austral spring [26]; meaning that when spring arrives, it could be in one of IOD phases. Thus, for the purpose of predicting spring rainfall

(Sep–Nov), the authors decided to take Dec_n-1-Aug_n values of DMI to bring into account as much information related to IOD as possible. In this way, a broader range of monthly DMI values are explored (nine months lags). The output of ANN would be the spring rainfall at year "n".

Early stop technique was used to stop the network from over fitting. Number of hidden neurons was chosen based on trial and error considering 5, 10, 15, 20, 25, 30, 35 hidden nodes. In this study, 1900–2006 was selected as the training and validation period and 2007–2009 was used as test period. The data were normalized between the range of 1 and 0 with Eq. (40.4). The models were evaluated using mean square error (MSE), Pearson correlation (R) and index of agreement (d) (Eq. 40.5).

$$X_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \tag{40.4}$$

$$d = 1 - \left(\frac{\left[\sum |\hat{y}_i - X_i|^2 \right]}{\left[\sum (|\hat{y}_i - \bar{x}_i| + |X_i - \bar{x}_i|)^2 \right]} \right) \tag{40.5}$$

where $\hat{y}_i$ is the predicted value of the ith observation and X_i is the normalized ith observation.

40.4 Results and Discussions

ANN modeling was carried out for separate and combined climate indices considering 5, 10, 15, 20, 25, 30 and 35 hidden neurons for three regions in Victoria. When performing an ANN modeling, the best model should be selected based on the error of the validation set; the best model of each input set was chosen based on the least MSE of the validation set. The best hidden neuron varied for different models and a unity was not observed among the models. Tables 40.1, 40.2, 40.3 show the MSE of the models for the three stations. The models are named after their input set. The best models are italicized in each table.

It can be seen from Table 40.1 that the models which have been developed for Horsham show same validation error. Among the single index models, ENSO indicators, i.e. NINO3, NINO4, NINO3.4 and SOI show an MSE of 0.04 while DMI based model has a very lower MSE of 0.015. These results demonstrate that comparing lagged ENSO and IOD indicators, IOD based model is proving to have less error compared to ENSO models. Combining the lagged ENSO and IOD indicators improved the performance of the ENSO based models from 0.04 to 0.03, however single IOD lagged based model with an MSE of 0.015 is by far the best model for Horsham (Table 40.1).

Models being developed for Melbourne (central Victoria) show similar results as of Horsham. It can be seen from Table 40.2 that all single index models are

Table 40.1 MSE of the best models for Horsham

Model	Train	Validation	Test
NINO3	0.03	0.04	0.06
NINO4	0.04	0.04	0.09
NINO3.4	0.03	0.04	0.07
SOI	0.03	0.04	0.04
DMI	*0.02*	*0.015*	*0.004*
NINO3-DMI	0.02	0.03	0.04
NINO4-DMI	0.01	0.03	0.10
NINO3.4-DMI	0.02	0.03	0.05
SOI-DMI	0.03	0.03	0.03

Table 40.2 MSE of the best models for Melbourne

Model	Train	Validation	Test
NINO3	0.03	0.03	0.05
NINO4	0.03	0.03	0.04
NINO3.4	0.07	0.03	0.03
SOI	0.01	0.03	0.05
DMI	*0.02*	*0.03*	*0.005*
NINO3-DMI	0.03	0.03	0.02
NINO4-DMI	0.02	0.04	0.03
NINO3.4-DMI	0.02	0.03	0.02
SOI-DMI	0.05	0.04	0.04

showing the same error in the validation set (MSE = 0.03); two of the combined indices models (NINO3-DMI and NINO3.4-DMI) are having the same error while for NINO4-DMI and SOI-DMI the MSE has increased to 0.04. This shows that combining lagged ENSO and IOD does not improve the performance of the models for this region and in some cases increases the error. Among the models with less MSE (0.03) the model with a lower error in the testing set i.e. DMI-based model, was chosen as the best model for this station; In this way the model with a higher generalization ability was chosen with an MSE of 0.005 for the testing set.

Table 40.3 shows the results of models being developed for Orbost. It can be seen from this table that MSE for DMI and SOI-DMI based models is 0.04 while for the rest of the models it is 0.05. It is clear that combining lagged ENSO_DMI indicators did not improve the results of forecast except for SOI-DMI model. These two models, i.e. DMI and SOI-DMI have a generalization error of 0.02; thus, DMI and SOI-DMI is considered the best models for this station. The results of these two models will be further discussed.

Combining ENSO and IOD did not improve the model performances. One explanation is that ENSO and IOD have different effective periods on Victorian rainfall; using the same lagged months for both indices in order to predict rainfall might not capture their nature efficiently. Thus, further study is underway to investigate the use of different lags of ENSO and IOD for rainfall prediction.

Table 40.3 MSE of the best models for Orbost

Model	Train	Validation	Test
NINO3	0.02	0.05	0.06
NINO4	0.05	0.05	0.02
NINO3.4	0.06	0.05	0.03
SOI	0.02	0.05	0.04
DMI	*0.04*	*0.04*	*0.02*
NINO3-DMI	0.08	0.05	0.007
NINO4-DMI	0.006	0.05	0.19
NINO3.4-DMI	0.05	0.05	0.01
SOI-DMI	*0.04*	*0.04*	*0.02*

Pearson correlation of these models with the rainfall of the three stations is shown in Table 40.4. By using lagged predictors as inputs in the ANN modeling, the authors were able to predict spring rainfall 3 consecutive years in advance for Horsham, Melbourne and Orbost. It can be seen from Table 40.4 that correlation coefficient for the testing set of Horsham and Melbourne is 0.99 and 0.98 respectively, while for Orbost, even though the two best models have the same MSE in the validation and testing set, however Table 40.4 shows that the correlation of SOI-DMI based model for the testing set is better than the DMI model. For better assessment of the model performance, an additional criterion, the Index of agreement (d) [27] has been chosen for model comparison. A 'd' value close to 1 indicates a better fitted model. Table 40.5 shows 'd' values for the three stations. Horsham is having a high 'd' value both in validation and testing set while for Melbourne the test set has a higher 'd' value. The 'd' value for Orbost for the SOI-DMI model is nearly 0.50 confirming that this model is a better model than DMI model for this region.

Figures 40.3, 40.4 and 40.5 shows the best models for the three stations. It can be seen from these figures that observed spring rainfall of these regions follow a very noisy pattern. Using the lagged climate indices as predictors, ANN was able

Table 40.4 Pearson correlation of the best models

Model		Train	Validation	Test
Horsham		0.66	0.74	0.99
Melbourne		0.56	0.34	0.98
Orbost	DMI	0.34	0.52	−0.71
	SOI-DMI	0.27	0.50	0.30

Table 40.5 'd' values for the best models

Model		Train	Validation	Test
Horsham		0.75	0.81	0.94
Melbourne		0.62	0.37	0.92
Orbost	DMI	0.50	0.58	0.14
	SOI-DMI	0.49	0.66	0.47

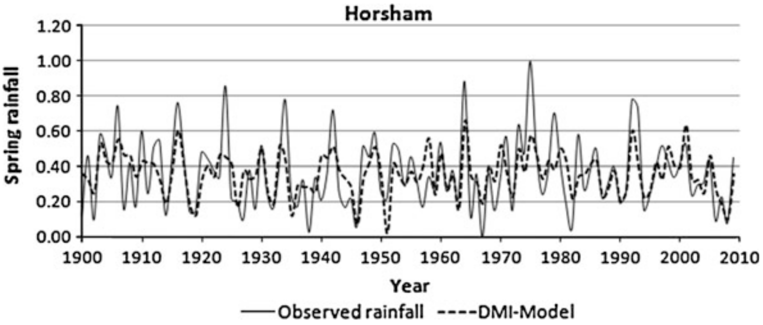


Fig. 40.3 ANN model for spring rainfall for Horsham (west Victoria), (1900–1990: training period, 1991–2006: validation period, 2007–2009: testing period)

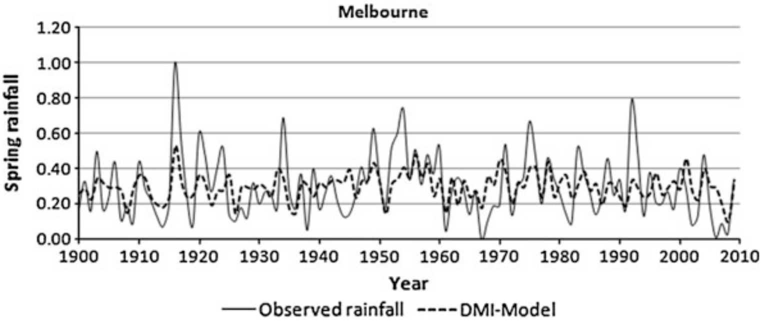


Fig. 40.4 ANN model for spring rainfall for Melbourne (central Victoria), (1900–1990: training period, 1991–2006: validation period, 2007–2009: testing period)

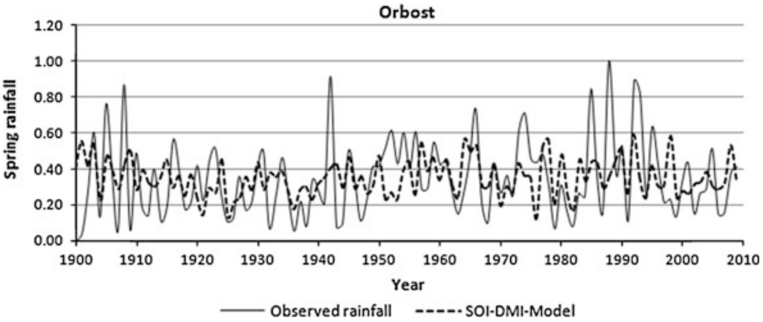


Fig. 40.5 ANN model for spring rainfall for Orbost (east Victoria), (1900–1990: training period, 1991–2006: validation period, 2007–2009: testing period).)

to model the observed rainfall in a way that not only the models follow the pattern of rainfall during several years but also they are able to predict long-term future; ANN is smoothly fitting the series capturing all the peaks and minimum, however there seems to be an underestimation in the models which can be improved by using a k-fold cross validation in future work.

40.5 Conclusions and Future Work

This study attempted to predict spring rainfall three consecutive years in advance by considering single and combined lagged climate indices as potential predictors. A nonlinear Artificial Neural Networks method was performed in order to investigate the predictability of spring rainfall using lagged ENSO, IOD and ENSO-IOD indicators. NINO3, NINO4, NINO3.4 and SOI were chosen as ENSO indicators and DMI was chosen as IOD indicator, the previous studies were focusing on finding the effect of these indices separately on Victorian rainfall but could not achieve a predictability of more than 30 %. This study discovered that the single lagged climate indices have more effect on rainfall predictability in west and central Victoria than the combined lagged climate indices. For east Victoria there seemed to be no difference in using single or combined climate indices and this region followed a different pattern than the other two regions; combined SOI-DMI model was chosen as the best model for east Victoria which allows for a smoother prediction regarding correlation of coefficient. Using climate indices in an ANN model increased the model correlation up to 0.99, 0.98 and 0.30 for the three case study stations of Horsham, Melbourne and Orbost respectively. It seems that both lagged ENSO and IOD have a strong effect on Victoria, however IOD has a higher effect on the centre and west of Victoria compared to ENSO, while ENSO and IOD both have a strong effect on east Victoria.

There is a need to further investigate this method on other different rainfall stations which will be covered in future studies. Also the effect of each lag on the predictability of rainfall needs further attention in a sensitivity analysis procedure.

References

1. Lau K, Weng H (2001) Coherent modes of global SST and summer rainfall over China: an assessment of the regional impacts of the 1997–98 El Nino. J Clim 14:1294–1308
2. Yufu G, Yan Z, Jia W (2002) Numerical simulation of the relationships between the 1998 Yangtze River valley floods and SST anomalies. Adv Atmos Sci 19:391–404
3. Barsugli JJ, Sardeshmukh PD (2002) Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin. J Clim 15:3427–3442
4. Hartmann H, Becker S, King L (2008) Predicting summer rainfall in the Yangtze River basin with neural networks. Int J Climatol 28:925–936

5. Chattopadhyay G, Chattopadhyay S, Jain R (2010) Multivariate forecast of winter monsoon rainfall in India using SST anomaly as a predictor: neurocomputing and statistical approaches. CR Geosci 342:755–765
6. Shukla RP, Tripathi KC, Pandey AC, Das IML (2011) Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach. Atmos Res 102:99–109
7. Kirono DGC, Chiew FHS, Kent DM (2010) Identification of best predictors for forecasting seasonal rainfall and runoff in Australia. Hydrol Process 24:1237–1247
8. Risbey JS, Pook MJ, McIntosh PC, Wheeler MC, Hendon HH (2009) On the remote drivers of rainfall variability in Australia. Mon Weather Rev 137:3233–3253
9. Saji NH, Goswami BN, Vinayachandran PN, Yamagata T (1999) A dipole mode in the tropical Indian ocean. Nature 401:360–363
10. Meneghini B, Simmonds I, Smith IN (2007) Association between Australian rainfall and the Southern annular mode. Int J Climatol 27:109–121
11. Hendon HH, Thompson DWJ, Wheeler MC (2007) Australian rainfall and surface temperature variations associated with the Southern hemisphere annular mode. J Clim 20(11):2452–2467
12. Ummenhofer CC, Sen Gupta A, Pook MJ, England MH (2008) Anomalous rainfall over southwest Western Australia forced by Indian Ocean sea surface temperatures. J Clim 21:5113–5134
13. England MH, Ummenhofer CC, Santoso A (2006) Interannual rainfall extremes over southwest Western Australia linked to Indian Ocean climate variability. J Clim 19:1948–1969
14. Evans AD, Bennett JM, Ewenz CM (2009) South Australian rainfall variability and climate extremes. Clim Dyn 33:477–493
15. Nicholls N (2010) Local and remote causes of the southern Australian autumn-winter rainfall decline, 1958–2007. Clim Dyn 34:835–845
16. Verdon DC, Wyatt AM, Kiem AS, Franks SW (2004) Multidecadal variability of rainfall and streamflow: Eastern Australia. Water Resour Res 40:W10201
17. Murphy BF, Timbal B (2008) A review of recent climate variability and climate change in Southeastern Australia. Int J Climatol 28:859–879
18. Verdon-Kidd DC, Kiem AS (2009) On the relationship between large-scale climate modes and regional synoptic patterns that drive Victorian rainfall. Hydrol Earth Syst Sci 13:467–479
19. Kiem AS, Verdon-Kidd DC (2009) Climatic drivers of Victorian streamflow: is ENSO the dominant influence. Aust J Water Resour 13:17–29
20. Schepen A, Wang QJ, Robertson D (2012) Evidence for using lagged climate indices to forecast Australian seasonal rainfall. J Clim 25(4):1230–1246
21. Abbot J, Marohasy J (2012) Application of artificial neural networks to rainfall forecasting in Queensland. Aust Adv Atmos Sci 29(4):717–730
22. Mekanik F, Imteaz MA (2012) A multivariate artificial neural network approach for rainfall forecasting: case study of Victoria, Australia. In: Proceedings of the world congress on engineering and computer Science 2012, Lecture notes in engineering and computer science, 24–26 October, 2012, San Francisco, USA, pp 557–561
23. De Vos N, Rientjes T (2005) Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation. Hydrol Earth Syst Sci Discuss 2(1):365–415
24. Yilmaz AG, Imteaz MA, Jenkins G (2011) Catchment flow estimation using artificial neural networks in the mountainous Euphrates basin. J Hydrol 410(1–2):134–140
25. Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ Model Softw 15:101–124
26. Cai W, Van Rensch P, Cowan T, Hendon H (2011) Teleconnection pathways of ENSO and the IOD and the mechanisms for impacts on Australian rainfall. J Clim 24(15):3910–3923
27. Willmott CJ (1982) Some comments on the evaluation of model performance. Bull Am Meteorol Soc 63:1309–1369

Chapter 41
A Simplex-Crossover-Based Multi-Objective Evolutionary Algorithm

Claudio Comis Da Ronco and Ernesto Benini

Abstract The key issue for an efficient and reliable multi-objective evolutionary algorithm is the ability to converge to the True Pareto Front with the least number of objective function evaluations, while covering it as much as possible. To this purpose, in a previous paper performance comparisons showed that the Genetic Diversity Evolutionary Algorithm (GeDEA) was at the same level of the best state-of-the-art MOEAs due to it intrinsic ability to properly conjugate exploitation of current non-dominated solutions and the exploration of the search space. In this paper, an improved version, namely the GeDEA-II, is proposed which features a novel crossover operator, the Simplex-Crossover, and a novel mutation operator, the Shrink-Mutation. GeDEM operator was left unchanged and completed using the non-dominated-sorting based on crowding distance. The comparison among GeDEA-II and GeDEA, as well as with three other modern elitist methods, on different extremely multidimensional test problems, clearly indicates that the performance of GeDEA-II is, at least in these cases, superior. In addition, authors aimed at putting in evidence the very good performance of GeDEA-II even in extremely multidimensional landscapes. To do this, four test problems were considered, and the GeDEA-II performance tested as the number of decision variables was increased. In particular, *ZDT* test functions featured a number of decision variables ranging from the original proposed number up to 1,000, whereas on *DTLZ* the decision variables were increased up to 100 times the original proposed number. Results obtained contribute to demonstrate further the GeDEA-II breakthrough performance.

C. Comis Da Ronco
HIT09 S.r.l, Galleria Storione 8, 35131 Padova, Italy
e-mail: c.comis@hit09.com

E. Benini (✉)
Department of Industrial Engineering, University of Padova,
Via Venezia 1, 35131 Padova, Italy
e-mail: ernesto.benini@unipd.it

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_41,

Keywords Empirical–comparison · Evolutionary algorithms · Genetic diversity · Multi objective optimization · Pareto optimality · Simplex crossover · Shrink mutation

41.1 Introduction

In the past, several powerful Multi-Objective Evolutionary Algorithms (MOEAs) were proposed, e.g. NSGA-II [8], SPEA-II [20] and IBEA [19]. GeDEA [16] algorithm, which was designed around the genetic diversity preservation mechanism called GeDEM, proved to be able to compete and, in some cases, to outperform, the aforementioned MOEAs as far as speed of convergence and covering uniformity of the Pareto Front are concerned. However, the common drawback of all of the previously mentioned multi-objective evolutionary algorithms concerns the huge amount of objective function evaluations (or number of generations) required to reach and sufficiently cover the Pareto Front.

To try to overcome this common weakness, during the last decade several authors started hybridizing evolutionary algorithms (EAs) with local search (LS) operators.

Several examples can be found in literature (some recent works are presented in [12] and [13]).

In spite of the different frameworks, in all the previously mentioned works, the local search, based on the Simplex algorithm, and the global exploration based on the Evolutionary algorithm, are performed separately, in a sequential manner, that is, a point of the search space is calculated via either the first or the latter.

In the authors' opinion, the previously mentioned examples of hybridization with local search often degrade the global search ability of MOEAs. Moreover, local search based on the Nelder and Mead requires additional and several functions evaluations.

In this paper, GeDEA-II is presented, aiming at reducing the potential weaknesses of its predecessor and competitors, while retaining its very good performance, that is, a good balance between exploration and exploitation. In particular, we propose a different approach to combine the Evolutionary algorithm-based global search and the Simplex theory, since global exploration and local search are intimately related and performed simultaneously, in such a way that they take advantage from each other. In details, we introduce a novel crossover operator, which we called the "Simplex-crossover" and which will be described hereafter; following this, the individuals created by the proposed algorithm via the Simplex-based crossover undergo mutation in a subsequently step, using another new typer of operator, which we called the "Shrink-Mutation", so as to promote global search capabilities of the algorithm. Moreover, important modifications have been brought about to the original Simplex theory, in order to enhance further the local search capabilities without penalizing the exploration of the search space.

The main differences of GeDEA-II in comparison with GeDEA regard its new Simplex-Crossover operator, and its new Shrink-Mutation operator. The diversity preserving mechanism, the Genetic Diversity Evaluation Method (GeDEM) already used in the GeDEA release, was retained in GeDEA-II and left unchanged due to its superior performance compared to other types of mechanisms.

41.2 Genetic Diversity Evolutionary Algorithm (GeDEA)

The Genetic Diversity Evolutionary Algorithm II (GeDEA-II), is a framework that is strictly designed around GeDEM [16] to exalt its characteristics. To briefly introduce the GeDEM principle, it is worth underlining that the multi-objective optimization process has two objectives, which are themselves conflicting: the convergence to the Pareto-optimal set and the maintenance of genetic diversity within the population. The basic idea of GeDEM is to actually use these objectives during the evaluation phase and to rank the solutions with respect to them, emphasizing the non-dominated solutions as well as the most genetically different. To this purpose, GeDEM computes the actual ranks of the solutions maximizing (1) the ranks scored with respect to the objectives of the original MOOP, the non-dominated solutions having the highest rank, and (2) the values assigned to each individual as a measure of its genetic diversity, calculated according to the chosen distance metric, i.e. the (normalized) Euclidean distance in the objective functions space.

41.3 Genetic Diversity Evolutionary Algorithm-II (GeDEA-II)

GeDEA proved to be an efficient algorithm, able to explore widely the search space, while exploiting the relationships among the solutions. In order to enhance GeDEA algorithm performance further, several main features were added to the previous GeDEA version, yet retaining its constitutive framework. The main innovation is the novel crossover operator, namely the Simplex-crossover, which substitutes the previous Uniform crossover. A novel mutation operator was also developed, namely the Shrink-mutation, which allows exploring more effectively the design space. The remaining steps characterizing GeDEA algorithm, in particular the GeDEM, were left unchanged. The latter was integrated with the Non-Dominating sorting procedure based on the crowding distance, developed and thoroughly described in [8]. The following sections present a detailed overview of the work already described in [6].

41.3.1 The SIMPLEX Crossover

In [17], authors proposed a simplex crossover (SPX), a new multi-parent recombination operator for real-coded GAs. The experimental results with test functions used in their studies showed SPX well performed on functions having multimodality and/or epistasis. However, the authors did not consider the application of the SPX to multiobjective problems. Moreover, they did not consider the possibility to take into account the fitness of the objective function/s as the driving force of the simplex. Therefore, we decided to integrate in the GeDEA-II the simplex crossover with these and further new distinctive features. Unlike the Simplex-crossover presented in [17], in GeDEA-II only two parents are required to form a new child. These two parents are selected according to the selection procedure from the previous population, and combined following the guidelines of the simplex algorithm. Let assume ***p1***, ***p2*** being the two parent vectors, characterized by different, multiple fitness values, the child vector ***Child*** is formed according to the reflection move described in [15]:

$$\mathbf{Child} := (1 + Refl) \cdot \mathbf{M} - Refl \cdot \mathbf{p}_2 \tag{41.1}$$

where ***Child*** is the new formed child and *Refl* is the reflection coefficient. It is assumed that ***p1*** is the best fitness individual among the two chosen to form the ***Child***, whereas ***p2*** the worst one. *Refl* coefficient is set equal to a random number ($refl \in [0, 1]$), unlike the elemental Simplex theory, which assumes a value equal to 1 for the *Refl* coefficient. This choice allows to create a child every time distant in a random manner from the parents, hence to explore more deeply the design space. Since the Simplex algorithm is itself a single-objective optimizer, a strategy was implemented to adapt it to a multi-objective algorithm: the objective function considered to form the new child is chosen randomly in order to enhance the design space exploration of the crossover, required in highly dimensional objective spaces.

This new crossover operator was expected to combine both exploration and exploitation characteristics. In fact, the new formed child explores a design space region opposite to that covered by the worst parent, that means it explores a region potentially not covered so far. In the early stages of the evolution, this means that child moves away from regions covered from bad parents, while exploring new promising ones. In addition, the characteristics of the good parents are deeply exploited to accelerate the evolution process.

During evolution, GeDEA-II makes use exclusively of the Simplex Crossover until three-quarters of the generations has been reached. After that, Simplex Crossover is used alternatively with the Simulated Binary Crossover (SBX) (described for the first time in [1]) with a switching probability of 50 percent.

41.3.2 The Shrink Mutation

As far as mutation is concerned, a new Shrink-mutation operator is introduced in the GeDEA-II.

In the literature, this kind of mutation strategy is referred to as *Gaussian mutation* [3], and conventional implementations of Evolutionary Programming (EP) and Evolution Strategies (ES) for continuous parameter optimization using Gaussian mutations to generate offspring are presented in [2] and [10], respectively.

In general, mutation operator specifies how the genetic algorithm makes small random changes in the individuals in the population to create mutation children. Mutation provides genetic diversity and enables the genetic algorithm to search a broader space. Unlike the previous version of mutation featuring GeDEA algorithm, where some bits of the offspring were randomly mutated with a probability p_{mut}, here the mutation operator adds a random number taken from a Gaussian distribution with mean equal to the original value of each decision variable characterizing the entry parent vector. The shrinking schedule employed is:

$$Shrink_i := Shrink_{i-1} \cdot \left(1 - \frac{ignr}{ngnr}\right) \tag{41.2}$$

where $Shrink_i$ is a vector representing the current mutation range allowed for that particular design variable, *ignr* represents the current generation and *ngnr* the total number of generations. The shape of the shrinking curve was decided after several experimental tests. The fact that the variation is zero at the last generation is also a key feature of this mutation operator. Being conceived in this manner, the mutation allows to deeply explore the design space during the first part of the optimization, while exploiting the non-dominated solutions during the last generations. Once the current variation range has been calculated, one decision variable of a selected child is randomly selected, and mutated according to the following formula:

$$Child_{mut} := Child_{cross} + [Shrink_i] \tag{41.3}$$

Unlike crossover operator, which generates all the offspring, mutation is applied only on a selected part of the offspring. Before starting offspring mutation, offspring

Table 41.1 Original and proposed number of generations for the *ZDT* and *DTLZ* test problems

	Number of generations	
	Original version problems	Proposed test problems
ZDT3	250	40
ZDT6	250	30
DTLZ3	500	150
DTLZ7	200	100

population is randomly shuffled to prevent locality effects. After that, a pre-established percentage (fixed to 40 % for all of the test problems) of the individuals are selected for mutation. The initial Shrink factor is set equal to the whole variation range of the design variables. This mutation operator was found to be powerful especially in multi-objective problems requiring a huge exploration of the design space.

41.4 Comparison with Other Multiobjective Evolutionary Algorithms

In order to judge the performance of the GeDEA-II, a comparison with other different state-of-the-art multi-objective EAs was performed. SPEA-2 [20], NSGA-II [8] and IBEA [19] were chosen as competitors, and their performance against GeDEA-II was measured on two test problems featuring the characteristics that may cause difficulties in converging to the Pareto-optimal front and in maintaining diversity within the population [7]: discrete Pareto fronts, and biased search spaces. In addition, their performance was tested also on two more recent and more challenging benchmark test functions chosen among the scalable Test Problems presented in [9]. The four test functions, the methodology and the metric of performance used in the comparison are briefly recalled in the following for easy reference.

41.4.1 Test Functions

Here only four test problems are presented due to layout constraints. The original version of ZDT_3 and ZDT_6 presented in [18] featured 30 and 10 decision variables, respectively. Here we propose them with 100 decision variables. As regards $DTLZ_3$, the number of variables suggested in [9] is 12. Here we propose it with 22 decision variables, respectively. As regards $DTLZ_7$, we increased the number of decision variables from the original one equal to 22, up to 100.

41.4.2 Methodology

The methodology used in [18] is strictly followed. GeDEA and competitors are executed 30 times on each test function. There are different parameters associated with the various algorithms, some common to all and some specific to a particular one. In order to make a fair comparison among all the algorithms, most of these constants are kept the same. In GeDEA-II, GeDEA and in competitors algorithms, the population size is set to 100. In the following, the parameters of the competitors MOEA are reported following the terminology used in PISA implementation. The individual mutation probability is always 1 and the variable mutation probability is fixed at $1/n$, n being the number of the decision variables of the test problem considered. The individual recombination probability along with the variable recombination probability are set to 1. The variable swap probability is set to 0.5. $\eta_{mutation}$ is always set to 20 and $\eta_{recombination}$ is fixed to 15. For IBEA algorithm, tournament size is always set to 2, whereas additive epsilon is chosen as the indicator. Scaling factor *kappa* is set to 0.05, and *rho* factor is fixed to 1.1. For all of the competitors, tournament size is given a value equal to 2. NSGA-II, SPEA-2 and IBEA are run with the PISA[1] implementation [4], with exactly the same parameters and variation operators. The maximum number of generations for test functions ZDT_6 is set to 30 for all the algorithms and to 40 for test function ZDT_3. For test function $DTLZ_7$ the number of generations is set to 100, whereas to 150 for $DTLZ_3$ test function. In Table 41.1, the original number of generations characterizing test problems presented in [18] and [9], is compared to the ones used here. The number of generations was intentionally reduced in order to test the convergence properties of the investigated algorithms, and contribute to justify the different results reported here, when compared to those presented in the original papers [9, 18].

41.4.3 Metric of Performance

Different metrics can be defined to compare the performance of EAs with respect to the different goals of optimization itself [18]: how far is the resulting nondominated set from the Pareto front, how uniform is the distribution of the solutions along the Pareto Approximation set/front, how wide is the Pareto Approximation set/front. For measuring the quality of the results, we have employed the Hypervolume approach, due to its construction simplicity and for the reason, which will be soon explained. The hypervolume approach measures how much of the objective space is dominated by a given nondominated set. Zitzler et al. [14] state it as the most appropriate scalar indicator since it combines both the distance of solutions (towards some utopian trade-off surface) and the spread of solutions. The Hypervolume[2] is defined as the area of coverage of PF_{known} with respect to the objective space for a two-objective MOP. In this work we use the version implemented by Fonseca et al. and presented in [11].

[1] This software is available for public use at PISA website http://www.tik.ee.ethz.ch/pisa/.

[2] The Hypervolume is a Pareto compliant indicator as stated in [5].

41.4.4 Results of Comparison

As in Zitzler et al. [18], Figs. 41.1, 41.3 and 41.4 show an excerpt of the non-dominated fronts obtained by the EAs and the Pareto-optimal fronts (continuous curves). The points plotted are the non-dominated solutions extracted from the union set of the outcomes of the first five runs, the best and the worst one being discarded. The performance of GeDEA-II is also compared to that of the competitors according to the hypervolume metric as defined in [11]. The distribution of these values is shown using box plots in Figs. 41.2 and 41.5. On each box, the central line represents the median, the edges of the box are the 25th and 75th %, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually, with a Plus sign. Results are normalized with the best Hypervolume value coming from the union set of all of the runs, extended to all of the algorithms. For each test problem, the reference point is assumed equal for all of the algorithms, and equal to the maximum value for each objective function from the union of all of the output points.

In general, the experimental results show that GeDEA-II is able to converge towards the True Pareto-optimal front and to develop a widely and well distributed non-dominated set of solutions. The comparison with the other three best-performing MOEAs according to the *Hypervolume* metric proves that the performance of GeDEA-II is somewhat superior. Considering the specific features of the two *ZDT* test functions, GeDEA-II shows similar performance both on multi-front and biased Pareto-optimal fronts. NSGA-II, SPEA-2 and IBEA seem instead to have more difficulties with discreteness (test function ZDT_3). The performance of GeDEA-II is particularly remarkable in the case of biased search space (test function ZDT_6) where it is also able to evolve a well-distributed non-dominated set. These results gain even more significance, since the number of decision variables was set to 100, unlike the original values of 30 (10 for the test function ZDT_6).

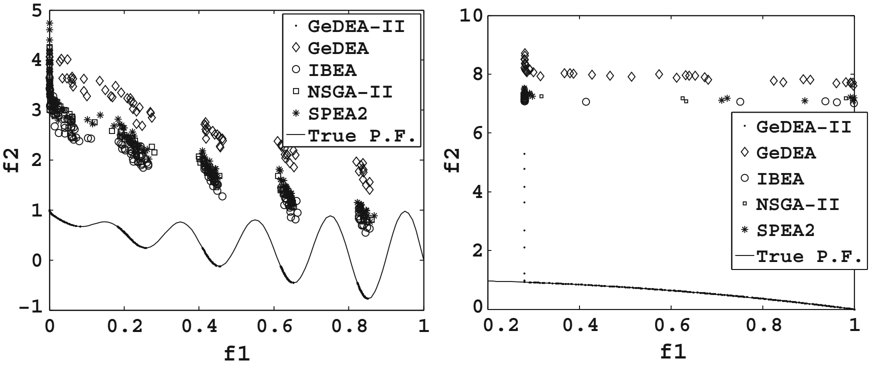


Fig. 41.1 Test functions ZDT_3 (*at the top*) and ZDT_6 (*at the bottom*)

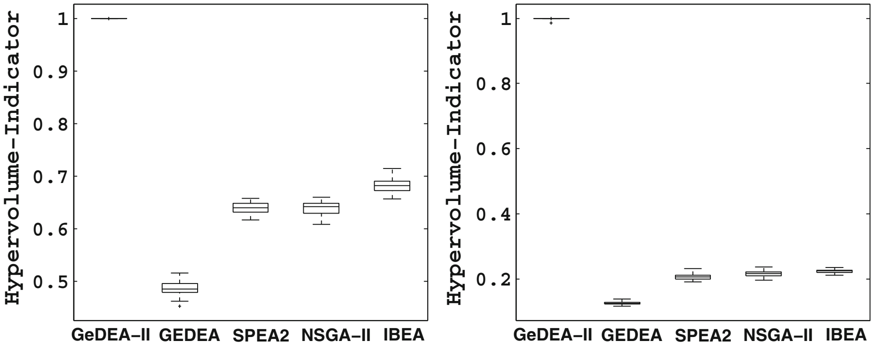


Fig. 41.2 Box plots based on the *Hypervolume* metric. Each square contains six box plots representing the distribution of *Hypervolume* values for the six algorithms. Results refer to the ZDT_3 (*at the top*) and ZDT_6 (*at the bottom*) test functions

As far as $DTLZ_3$ and $DTLZ_7$ test functions is concerned, GeDEA-II is able to reach the True Pareto Front, whereas the competitors remain trapped in the local Pareto Approximation Sets, as shown in Figs. 41.3 and 41.4.

Finally, box plots prove, in general, that the performance of GeDEA-II is superior to those of the competitors also as far as the repeatability of the results is concerned.

41.5 GeDEA-II Performance on Extremly Multidimensional Landscapes

In this section, authors aim at putting in evidence the outstanding performance of GeDEA-II even on high multidimensional environments. To do this, two test problems, chosen among those presented in Sect. 41.4.1 are considered, and the GeDEA-II performance tested by changing every time the number of decision variables. Test functions chosen for this test are the ZDT_4 and $DTLZ_3$, that is, the most difficult to solve problems, as stated in [18] and [9].

In Table 41.2, the number of variables and generations characterizing these tests are reported. In particular, ZDT_4 test function feature a maximum number of decision variables of 1000, whereas on $DTLZ_3$ test functions the maximum number of decision variables is increased up to 100 times the original proposed number [9].

To the best of the authors' knowledge, this is the first time a MOEA is tested on these test problems, with these number of decision variables.

For each test problems, we performed 30 independent runs for each number of decision variables, and the boxplots were then built, following the guidelines already given in Sect. 41.4.4. Y-axes are scaled in such a way the best run is given a value equal to 1. In Fig. 41.6, the boxplots showing GeDEA-II performance are

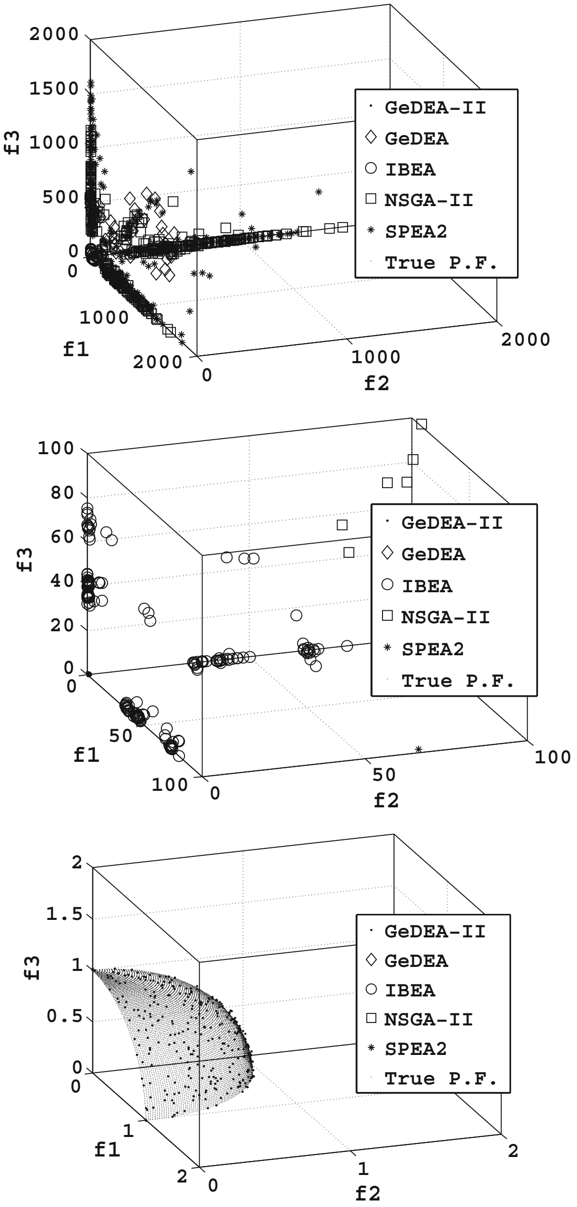


Fig. 41.3 Test function *DTLZ3*. From the left, Auto scale axes, medium zoom and true pareto front region

presented, as the decision variables are increased from the minimum value up to the maximum one.

Results clearly states that GeDEA-II performance is high-level. In each test problem, performance is never lower than 99% of the maximum value, no matter how many the decision variables are. This clearly demonstrate GeDEA-II manages

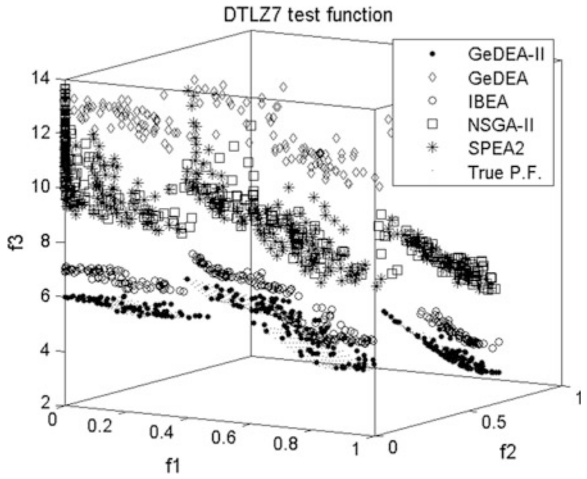


Fig. 41.4 Test function $DTLZ_7$

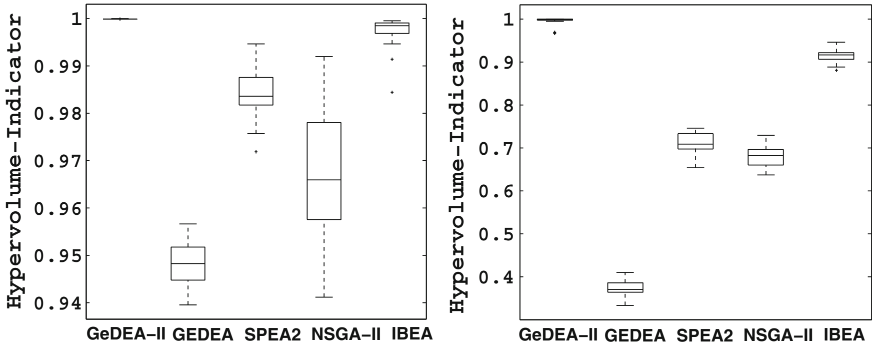


Fig. 41.5 Box plots based on the *Hypervolume* metric. Each square contains six box plots representing the distribution of *Hypervolume* values for the six algorithms. Results refer to the $DTLZ_1$ (*at the top*) and $DTLZ_7$ (*at the bottom*) test functions

Table 41.2 Minimum and maximum number of decision variables for the ZDT_4 and $DTLZ_3$ test problems

	Number of generations	Minimum number of decision variables	Maximum number of decision variables
ZDT4	40	10	1,000
DTLZ3	80	12	1,200

to evolve the initial population near to the True Pareto front, even when the number of decision variables is dramatically increased. Figure 41.7 shows in the objective space, the distribution of the final solutions obtained in the run with

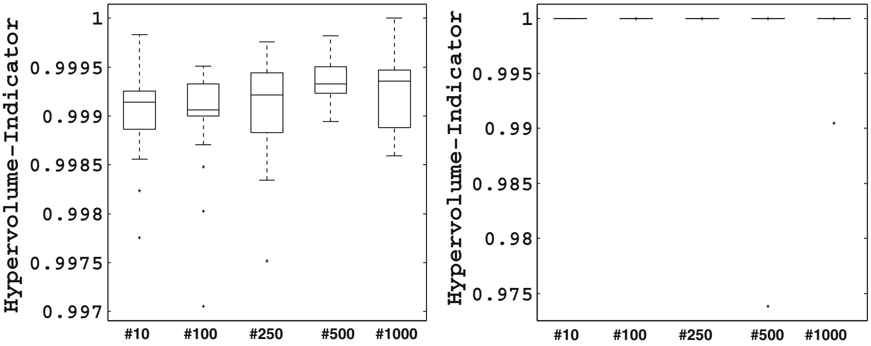


Fig. 41.6 Box plots based on the *Hypervolume* metric. Each square contains five box plots representing the distribution of *Hypervolume* values for the six number of decision variables. Results refer to the ZDT_4 (*at the top*) and $DTLZ_3$ (*at the bottom*) test functions

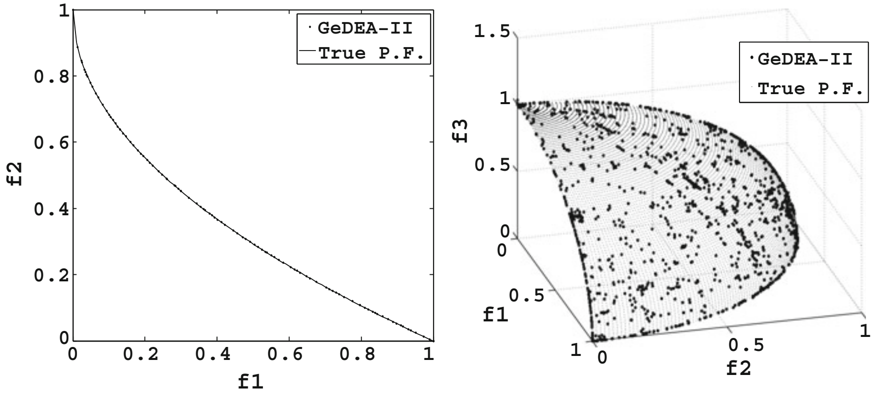


Fig. 41.7 Final approximation set reached by the GeDEA-II on test function ZDT_4 (*at the top*) and the non dominated solutions found by GeDEA-II on $DTLZ_3$ (*at the bottom*), featuring 1,200 decision variables

the lowest Hypervolume-value by the GeDEA-II for each test instance, for the maximum number of decision variables. It is evident that as regards the convergence to the True Pareto Front and spread of solutions, GeDEA-II performance is high level.

Appendix

Each of the three *ZDT* test functions, namely ZDT_3, ZDT_4 and ZDT_6 introduced in [18] is a two-objective minimization problem that involves a distinct feature among those identified in [7]. All the test functions are constructed in the same way, according to the guidelines in [7]:

$$\begin{aligned} Minimize &: T(x) = (f_1(x_1), f_2(x)) \\ subject\ to &: f_2(x) = g(x_2; \ldots; x_m) h(f_1(x_1), g(x_2; \ldots; x_m)) \\ where &: x = (x_1, \ldots, x_M) \end{aligned} \tag{A.41.1}$$

Function f controls vector representation uniformity along the Pareto approximation set. Function g controls the resulting MOP characteristics (whether it is multifrontal or has an isolated optimum). Function h controls the resulting Pareto front characteristics (e.g., convex, disconnected, etc.) These functions respectively influence search along and towards the true Pareto front, and the shape of a Pareto front in R^2. Deb [7] implies that a MOEA has difficulty finding PF_{true} because it gets "trapped" in the local optimum, namely PF_{local}. Test functions reported in this work feature an increased number of decision variables, when compared to their original versions reported in [18]. This choice was motivated by the authors' will of testing exploration capabilities of the algorithms also on highly dimensional test problems, and contributes to justify the results presented in Sect. 41.4.4.

- Test function ZDT_3 features a Pareto-optimal front disconnected, consisting of several noncontiguous convex parts:

$$\begin{aligned} f_1(x_1) &= (x_1) \\ g(x_2; \ldots; x_n) &= 1 + 9 \cdot \sum_{i=2}^{n} \frac{x_i}{(n-1)} \\ h(f_1, g) &= 1 - \left(\sqrt{\frac{f_1}{g}}\right) - \left(\frac{f_1}{g}\right) \cdot \sin(10\pi f_1) \end{aligned} \tag{A.41.2}$$

where $n = 100$ and $x_i \in [0,1]$. The Pareto-optimal front corresponds to $g(\mathrm{x}) = 1$.

The original version presented in [18] featured 30 decision variables.

- Test function ZDT_4 contains 21^9 local Pareto-optimal fronts and, therefore, tests for the EA ability to deal with multifrontality:

$$\begin{aligned} f_1(x_1) &= (x_1) \\ g(x_2; \ldots; x_n) &= 1 + 10(n-1) \cdot \sum_{i=2}^{n} (x_i{}^2 - 10\cos(4\pi x_i)) \\ h(f_1, g) &= 1 - \sqrt{\frac{f_1}{g}} \end{aligned} \tag{A.41.3}$$

where $n = 100$ and $x_i \in [0,1]$. The Pareto-optimal front is convex and corresponds to $g(\mathrm{x}) = 1$. The original version presented in [18] featured 10 decision variables.

- Test function ZDT_6 features two difficulties caused by the non-uniformity of the search space: first, the Pareto optimal solutions are nonuniformly distributed along the PF$_{true}$ (the front is biased for solutions for which $f_1(x_1)$ is near one); and second, the density of the solutions is lowest near the PF$_{true}$ and highest away from the front:

$$\begin{aligned} f_1(x_1) &= 1 - \exp(-4x_1)\sin^6(6\pi x_1) \\ g(x_2;\ldots;x_n) &= 1 + 9 \cdot \left(\sum_{i=2}^{n} \frac{x_i}{(n-1)} \right)^{1/4} \\ h(f_1, g) &= 1 - \left(\frac{f_1}{g} \right)^2 \end{aligned} \tag{A.41.4}$$

where n = 100 and $x_i \in$ [0,1]. The Pareto-optimal front is non-convex and corresponds to g(x) = 1. The original version presented in [18] featured 10 decision variables.

cFinally, two of the tri-objective minimization test functions designed by Kalyanmoy Deb, Lothar Thiele, Marco Laumanns and Eckart Zitzler, and presented in [9], are considered, in order to demonstrate the GeDEA-II capabilities on more than two-objectives test problems. In the following, n identifies the number of decision variables, M the number of objective functions, and $k = |x_M| = n - M + 1$ the number of variables of the functional $g(x_M)$. The number of variables was always increased when compared to that suggested by the authors in [9], whereas the decision variables range was left unchanged. These features help clarifying the different results between those reported in Sect. 41.4.4 and the original ones [9].

- Test function $DTLZ_3$ is similar to test function $DTLZ_2$, except for the function g, which introduces ($3^k - 1$) local Pareto-optimal fronts, and only one global Pareto-optimal front.

$$\begin{aligned} f_1(x) &= (1 + g(x_M))\cos(x_1\pi/2)\cos(x_2\pi/2) \\ f_2(x) &= (1 + g(x_M))\cos(x_1\pi/2)\sin(x_2\pi/2) \\ f_3(x) &= (1 + g(x_M))\sin(x_1\pi/2) \\ \text{and } g &= 100\left[k + \sum_{x_i \in x_M} (x_i - 0.5)^2 - \cos(20\pi(x_i - 0.5)) \right] \end{aligned} \tag{A.41.5}$$

where n = 22 and $x_i \in$ [0,1]. The number of variables suggested in [9] is 12.

- Test function $DTLZ_7$ features 2^{M-1} disconnected local Pareto-optimal regions in the search space. It is chosen to test the MOEA ability in finding and maintain stable and distributed subpopulations in all four disconnected global Pareto-optimal regions.

$$
\begin{aligned}
f_1(x) &= x_1 \\
f_2(x) &= x_2 \\
f_3(x) &= (1 + g(x_M))h \\
g &= 1 + \frac{9}{k}\sum_{x_i \in x_M}(x_i) \\
\text{and } h &= M - \sum_{i=1}^{M-1}\left[\frac{f_i}{1+g}(+\sin((1+3\pi f_i))\right]
\end{aligned}
\tag{A.41.6}
$$

where n = 100 and $x_i \in$ [0,1]. Once again, the number of decision variables was dramatically increased when compared to the original one, suggested in [9] for this test problem, and equal to 22.

References

1. Bhusan Agrawal R, Deb K (1994) Simulated binary crossover for continuous search space. Complex Sys 9:115–148
2. Bäck T (1996) Evolutionary Algorithms in Theory and Practice. Oxford University Press, Oxford
3. Bäck T, Schwefel H-P (1993) An overview of evolutionary algorithms for parameter optimization. Evol Comput 1(1):1–23
4. Bleuler S, Laumanns M, Thiele L, Zitzler E (2002) PISA–a platform and programming language independent interface for search algorithms. Springer, NY
5. Coello Coello CA, Van Veldhuizen DA, Lamont GB (2002) Evolutionary algorithms for solving multi-objective problems. Kluwer Academic Publishers, New York
6. Comis Da Ronco C, Benini E (2012) Gedea-II: a novel evolutionary algorithm for multi-objective optimization problems based on the simplex crossover and the shrink mutation. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science (2012) WCECS 2012, 24–26 Oct 2012. USA, San Francisco, pp 1298–1303
7. Deb K (1999) Multi-objective genetic algorithms: problem difficulties and construction of test problems. Evol Comput 7(3):205–230
8. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6(2):182–197
9. Deb K, Thiele L, Laumanns M, Zitzler E (2001) Scalable test problems for evolutionary multi-objective optimization. Computer Engineering and Networks Laboratory (TIK), TIK-Technical Report No. 112, Swiss Federal Institute of Technology
10. Fogel DB (1995) Evolutionary computation: toward a new philosophy of machine intelligence. IEEE Press, Piscataway, NJ
11. Fonseca CM, Paquete L, López-Ibáñez M (2006) An improved dimension-sweep algorithm for the hypervolume indicator. In: IEEE congress on evolutionary computation, pp 1157–1163
12. Ghiasi Hossein, Pasini Damiano, Lessard Larry (2011) A non-dominated sorting hybrid algorithm for multi-objective optimization of engineering problems. Eng Optim 43(1):39–59
13. Koduru P, Dong Z, Das S, Welch S, Roe JL, Charbit E (2008) A multiobjective evolutionary-simplex hybrid approach for the optimization of differential equation models of gene networks. IEEE Trans Evol Comput 12(5):572–590

14. Laumanns M, Zitzler E, Thiele L (2001) On the effects of archiving, elitism, and density based selection in evolutionary multi-objective optimization. In: Proceedings of the first international conference on evolutionary multi-criterion optimization, Springer-Verlag, London, UK, pp 181–196
15. Nelder JM, Mead R (1965) A simplex method for function minimization. Comput J 7(4):308–313
16. Toffolo A, Benini E (2002) Genetic diversity as an objective in multi-objective evolutionary algorithms. Evol Comput 11(2):151–157
17. Tsutsui S, Yamamura M, Higuchi T (1999) Multi-parent recombination with simplex crossover in real coded genetic algorithms. In: Proceedings of the GECCO-99, pp 657–644
18. Zitzler E, Deb K, Thiele L (2000) Comparison of multiobjective evolutionary algorithms: empirical results. Evol Comput 8(2):173–195
19. Zitzler E, Künzli S et al (2004) Indicator-based selection in multiobjective search. In: Yao X (ed) Parallel problem solving from nature (PPSN VIII). Springer-Verlag, Berlin, Germany, pp 832–842
20. Zitzler E, Laumanns M, Thiele L (2001) SPEA2: improving the strength pareto evolutionary algorithm. In: Technical report 103, computer engineering and networks laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich, Gloriastrasse 35, CH-8092 Zurich, Switzerland, May 2001

Chapter 42
Efficient Operational Management of Enterprise File Server with File Size Distribution Model

Toshiko Matsumoto, Takashi Onoyama and Norihisa Komoda

Abstract Toward efficient operational management of enterprise file server, we propose an estimation method for relationship between file number and cumulative file size in descending order of file size based on a model for file size distribution. We develop the model by weighted summation of multiple log normal distribution based on AIC. File size data from technical and non-technical divisions of a company show that our model fits well with observed distribution, and that the estimated relationship can be utilized for cost-effective operational management of file server.

Keywords Akaike's information criterion · Enterprise file server · File size · Log normal distribution · Operational management · Tiered storage

42.1 Introduction

Recently enterprise file server has received more and more attention, because of growing trend of unstructured files. In addition, file server is expected to handle archive need for e-Discovery, and real-time backup for Business Continuity Management. From a viewpoint of cost effectiveness, several solutions have been provided: tiered storage technology, de-duplication, deleting unnecessary files,

T. Matsumoto (✉) · T. Onoyama
Hitachi Solutions, Ltd., 4-12-7, Higashishinagawa, Shinagawa-ku, Tokyo 140-0002, Japan
e-mail: toshiko.matsumoto.jz@hitachi-solutions.com

T. Onoyama
e-mail: takashi.onoyama.js@hitachi-solutions.com

N. Komoda
Osaka University, 2-1, Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: komoda@ist.osaka-u.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_42,

usage report of file server, and so on [1, 2]. Tiered storage technology contributes not only to cost-effectiveness of high performance storage system but also to reducing cost for real-time backup in Business Continuity Management by separating active and inactive files. De-duplication reduces the size of data volume in a fileserver and thus is able to shorten backup time. Deleting unnecessary files can improve cost-performance of file servers and leads to more desirable usage of files. Usage report enables file server administrators to develop a plan for introducing tired storage technology, for utilizing de-duplication, and for deleting unnecessary files. Toward theoretical evaluation of algorithms implemented in these solutions, several statistical characteristics of files are reported [3–10]. They measured frequencies of file access, file extension type distribution, and fraction of duplicated contents. Some of them also suggested similarities between file size distribution and Pareto or log normal distribution [4, 8]. However, quantitative estimation requires a survey where

1. system files are discriminated from user files,
2. files are stored in a shared file server for collaboration,
3. data are from industrial file servers, and
4. models are evaluated by statistically testing goodness of fit.

We have worked on model for file size distribution, which is one of the most fundamental statistical characteristics of files [11]. In this chapter, we propose a model as a weighted summation of multiple log normal distributions. Number of log normal distributions is decided based of Akaike's Information Criterion (AIC) to prevent over fitting [12]. We also describe expanded cases of operational management of file server where theoretical model of file size distribution is utilized. Finally, the model and the estimation are evaluated with actual data.

42.2 About Data

We use data of shared file servers from 24 divisions of a company: 11 are technical divisions such as research and development, and the other 13 are non-technical divisions such as sales, marketing, and management. Some divisions have tens of thousands of files and others have more than one-million files. Sum of file sizes are between several gigabytes to several hundreds of gigabytes.

Sizes of files in a shared file server of Research and Development division of a company are plotted in Fig. 42.1. X-axis and y-axis represents number and file size in descending order, respectively. This graph is equivalent to transposed cumulative probability distribution, and demonstrates that there are very few large files. All the other divisions show the same tendency. Figure 42.2 shows the same data of Fig. 42.1 and data of Pareto distribution [8] in logarithmic scale. Since Pareto distribution is proposed for file size which is larger than some threshold, linear regression line of logarithm value of file size upon logarithm value of file number is calculated with Finney's correction method [13] from plot with file size of

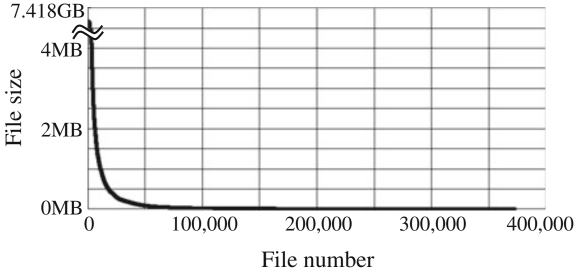


Fig. 42.1 Example plot of file size in descending order

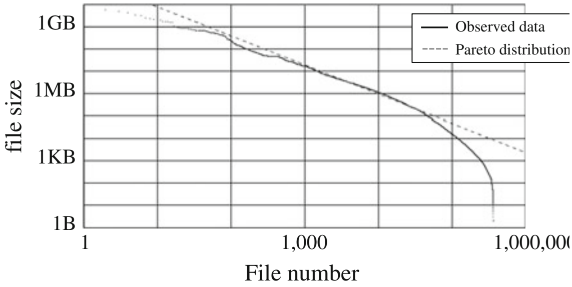


Fig. 42.2 Example plot of file size in descending order for observed data and Pareto distribution

100 KB or larger. In Fig. 42.2, Finney's correction has so small effect that linear regression line with and without correction almost completely overlaps.

Figure 42.3 shows file size histogram of the observed data and of two theoretical distributions proposed in previous studies [4, 8]. Number of classes of the histogram is decided based on Sturges' formula [14] shown in Eq. (42.1).

$$(\text{number of classes}) = 1 + \log_2 (\text{number of data}) \tag{42.1}$$

Log normal distribution is calculated by average and standard deviation of logarithm value of file size. Observed plot seems to have linearity in Fig. 42.2 and shows roughly bell-shaped form Fig. 42.3. However, frequency of observed data differs from Pareto distribution by more than 20,000 at 100 KB and from log normal distribution by almost 10,000 at peak in the histogram. Because of these apparent discrepancies, null hypotheses "Pareto distribution fits observed histogram" and "log normal distribution fits observed histogram" are rejected by χ^2 test with type I error rate of 0.01. Null hypothesis "there is not large amount of divergence between Pareto distribution and observed histogram" and "there is not large amount of divergence between log normal distribution and observed histogram" are also rejected by generalized χ^2 test [15] with divergence level of 0.25 % and type I error rate of 0.01. These null hypotheses are rejected even if file size threshold is set to 1 MB. In all other divisions, observed data differ significantly both from Pareto distribution and from log normal distribution. Therefore, our data demonstrates that there are statistically significant difference between these two theoretical models and observed data, even though they have roughly similar histogram form.

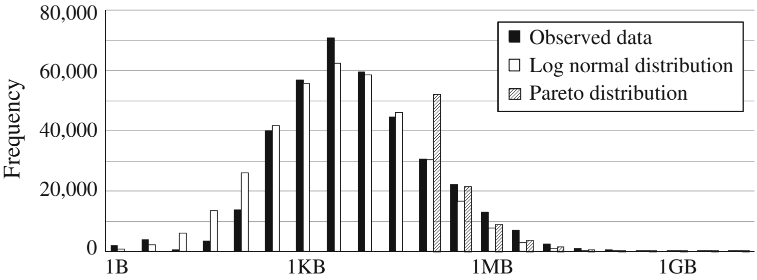


Fig. 42.3 Example plot of file size histogram for observed data and models in previous studies

42.3 File Size Modeling by Weighted Summation of Multiple Log Normal Distributions

42.3.1 AIC-Based Modeling

We propose file size modeling by weighted summation of multiple log normal distributions. Our model is based on observations where file size distribution depends on content type, such as movie files, database files, and executable files are larger than that of other files [3, 5], and based on an idea that more variety should be observed in general-purpose industrial file servers. Weighted summation of c log normal distributions are calculated to fit to observed data. Similarity is evaluated with χ^2 value from contingency table. χ^2 value is calculated for various c to minimize AIC. AIC is defined as Eq. (42.2), where L is maximum likelihood and k is number of degrees of freedom [12].

$$\mathrm{AIC} = 2\ln L + 2k \tag{42.2}$$

Likelihood ratio λ is L/L', where L and L' are maximum likelihood values under null hypothesis and in parameter space. AIC can be minimized by minimizing ($\chi^2 + 3c$), because of following three facts. First, L' can be treated as a constant value when observed data is given. Second, λ eventually closes to χ^2 distribution as sample size grows. Third, each one of log normal distribution adds three degrees of freedom: average, variance, and weight. We can prevent over fitting during model calculation, since AIC value is increased when too many parameters are adopted.

42.3.2 Comparison of Observed Data and Proposed Model of File Size

According to the previous section, theoretical distribution of logarithmic scaled file size is calculated as Eq. (42.3) where $N(\mu, \sigma^2)$ is normal distribution with average μ and variance σ^2.

$$
\begin{aligned}
&281960.8 \times N(3.28, 0.60)\\
&+ 42195.9 \times N(5.01, 0.10)\\
&+ 13121.3 \times N(5.95, 0.014)\\
&+ 12342.1 \times N(2.14, 0.028)\\
&+ 8620.5 \times N(4.23, 0.0081)\\
&+ 7144.9 \times N(6.50, 0.23)\\
&+ 3604.6 \times N(-0.01, 0.040)\\
&+ 2156.7 \times N(0.28, 0.0049)\\
&+ 844.5 \times N(1.38, 0.014)\\
&+ 252.4 \times N(7.88, 0.0025)\\
&+ 141.5 \times N(8.50, 0.090)\\
&+ 75.5 \times N(9.50, 9.00)
\end{aligned}
\qquad (42.3)
$$

Figure 42.4 shows file size histogram for observed data shown in Fig. 42.3 and for the proposed model.

Since theoretical distribution fits very well to the observed data at all classes, no statistically significant difference was observed. Equation (42.3) shows decrease of weights in an exponentially fashion. The decrease indicates that first term has a dominant effect to the distribution and supports that observed distribution shows roughly bell-shaped form. First term includes source code files and plain text files, that occupy large part of the files in the file server of the division. Second term includes HTML files and PDF files. HTML files have larger size than XML files do in average, because HTML files include specification documents of a programming language and files that are converted by a word processing software.

Observed distributions of file size in all the other 23 divisions fit very well to their corresponding models and demonstrate no statistically significant difference.

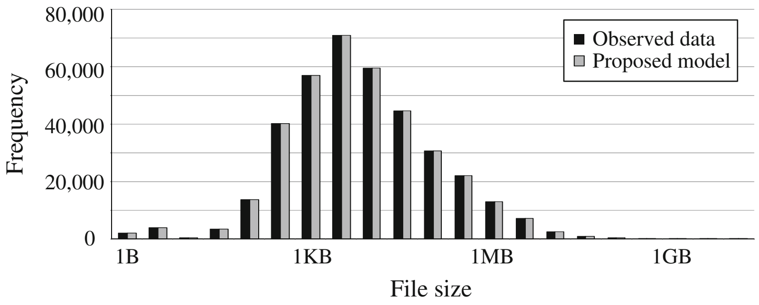


Fig. 42.4 Example plot of file size histogram for observed data and proposed model

42.4 Efficiency of Operational Management of File Server Based on Model of File Size Distribution

42.4.1 Cases of Operational Management Utilizing Model of File Size Distribution

File size distribution model can be directly utilized in the following three kinds of estimation for efficient operational management of file servers. First case is estimating effect of file moving policy in introducing process of tiered storage technology. Processing time of moving file depends both on size summation and on number of files. From these dependences, we can expect that assignment of a high priority to a large file achieves in an efficient moving policy: large volume is moved to lower cost storage in short time and that benefit of tiered storage technology can be realized efficiently with the policy. Since last access time and file size show no correlation (correlation coefficient <0.1 for all divisions in our data), file size distribution model in the Sect. 42.3.1 is expected to be effective to estimate relationship between number and size summation of files moved to lower cost storage.

Second case is estimating effect of approximate calculation in usage report of file server. For enterprise file servers, software products are provided for graph display of volume usage; such as sum of file size by file extension type or by value range of last access time. The display enables file server administrators to develop plans for expansion of file server capacity, for migration into a new file server, and for introducing tired storage technology. These software store metadata of files in database and query the database to obtain values for graph display. Since response time depends on number of files selected by the query, approximate calculation with summing up file size of a small number of large files can reduce processing time for usage report.

Third case is estimating effect of deleting unnecessary files by users. When millions of files are stored in a file server, it is obviously unrealistic to manually check deletability of all files one by one. Reduction of unnecessary files can be tractable only when large volume is deleted with manual confirmation of a small fraction of files. Deletability confirmation in descending order of file size is effective for efficient volume reduction when the files are deletable. When some files are confirmed to be undeletable, this confirmation policy is still effective for efficient estimation of upper limit of eventual reducible volume.

42.4.2 Estimating Efficiency Based on Cumulative File Size

When top n % of large files occupy p % of the total volume, relationship between n % and p % is equivalent to volume efficiency of introducing process of tiered

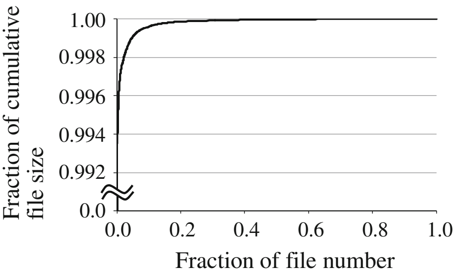


Fig. 42.5 Fraction of file number and of cumulative file size in descending order based on our model

storage technology where inactive files are moved to lower cost storage for the first case in the previous section. The relationship also represents efficiency of approximate calculation in usage report in the second case, and upper limit of reduction volume per confirmation number of file delectability in the third case. We can estimate the relationship by integrating theoretical distributions described in Sect. 42.3.1.

42.5 Experimental Result and Evaluation

42.5.1 Relationship Between Number of Files and Cumulative File Size

From Eq. (42.3) in Sect. 42.3, Fig. 42.5 shows relationship between fraction of file number n % and fraction of cumulative file size p % in descending order of file size. Value of p % rapidly reaches almost 100 %, whereas p % should be equal to n % in randomized order. Rapid increase of p % results from tiny fraction of large files as shown in Fig. 42.1. Since small value of n % can give large p % value, focusing on large files is expected to achieve efficient operational management in the all three cases of Sect. 42.4.1.

42.5.2 Accuracy Evaluation of Estimating Relationship Between Number and Cumulative Size of Files

In this section, we evaluate accuracy of our model and models in previous studies. As shown in Fig. 42.6, fraction of cumulative file size p % is calculated in three models for fraction of file number $n\ \% = 1\ \%, 5\ \%$, and 10 % in 24 divisions. p % is calculated from observed data for each value of n % in 24 divisions, and is compared with p % calculated in theoretical models to evaluate their accuracy. Figures 42.7, 42.8, and 42.9 show comparison results of observed data and estimated value by file size distribution models of Pareto distribution, log normal

Fig. 42.6 Procedures of accuracy evaluation of estimating relationship between number and cumulative size of files

```
for ($n$% = 1%, 5% and 10%) {
    for (1st, 2nd, …, 24th division) {
        calculate $p$% in Pareto distribution      ............. value1
        calculate $p$% in log-normal distribution      …… value2
        calculate $p$% in our model …………………… value3
        calculate $p$% from observed data    …………... value4
    }
}
compare value 1 and 4 as accuracy evaluation of estimating
  relationship between $n$% and $p$% with Pareto distribution
compare value 2 and 4 as accuracy evaluation of estimating
  relationship between $n$% and $p$% with log-normal distribution
compare value 3 and 4 as accuracy evaluation of estimating
  relationship between $n$% and $p$% with our model
```

Fig. 42.7 Fraction of cumulative file size occupied by top large files in observed data and in Pareto distribution

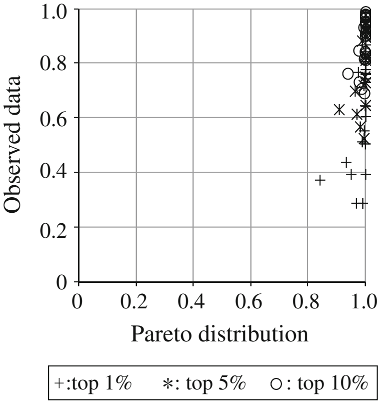
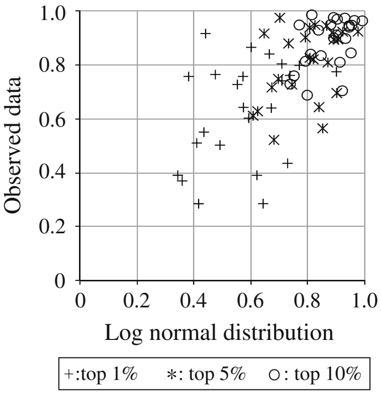


Fig. 42.8 Fraction of cumulative file size occupied by top large files in observed data and in log normal distribution



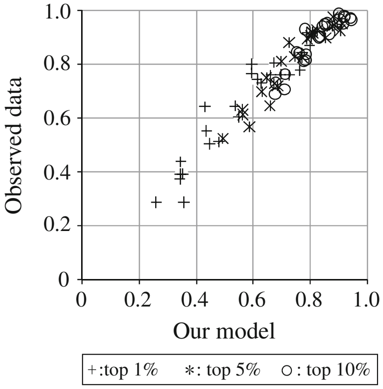


Fig. 42.9 Fraction of cumulative file size occupied by top large files in observed data and in our model

distribution, and our model calculated as described in Sect. 42.3.1 from 24 divisions. The plots show strong correlation between observed and estimated values of our model in Fig. 42.9 (correlation coefficient ≥ 0.9).

In contrast, no apparent correlation is observed for the case of Pareto distribution or log normal distribution (correlation coefficient ≤ 0.6). These results demonstrate that our model can estimate relationship between n % and p % more accurately than models of previous studies can, and is suitable for quantitative evaluation.

42.6 Shared File Servers in Technical/Non-Technical Divisions

We compare technical and non-technical divisions regarding fraction of cumulative file size p % for fraction of file number n % = 1 %, number of log normal distributions c calculated according to Sect. 42.3.1, and overview statistics of file server as shown in Table 42.1. T test with Bonferroni correction [16] reveals only

Table 42.1 Overview statistics of file server

No.	Overview statistics of file server
1	File number
2	File number (logarithmic value)
3	Sum of file sizes
4	Sum of file sizes (logarithmic value)
5	Average file size
6	Average file size (logarithmic value)
7	Maximum file size
8	Maximum file size (logarithmic value)
9	Maximum file size (logarithmic value)/file number (logarithmic value)

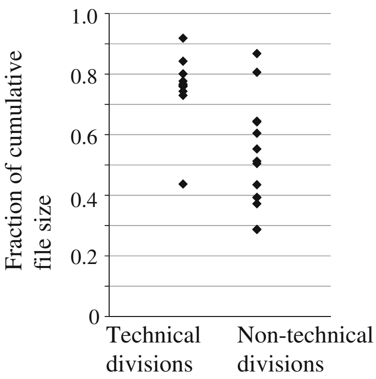


Fig. 42.10 Fraction of cumulative file size occupied by top 1 % large files in technical and non-technical divisions

p % shows statistically significant difference between technical and non-technical divisions. Values of *p* % in technical and non-technical divisions are shown in Fig. 42.10. These results mean that overview statistics of file server are not good at estimating *p* %, although *p* % depends on type of division. They also indicate that our model can estimate *p* % better.

42.7 Conclusion

In this study, we propose a file size distribution model in enterprise file server and describe that the model is effective for efficient operational management of file servers. Data of shared file servers from various technical and non-technical divisions demonstrate that file size distribution can be modeled as a weighted summation of multiple log normal distributions. The data also demonstrate that integrating the theoretical distribution gives accurate estimate for efficiency in three operational management cases: (1) introducing process of tired storage technology, (2) calculating usage report of file server, and (3) deleting unnecessary files. Our estimation can be applied to each file server by two conversions for each case. For the first case, cumulative file size is converted into cost effect on the basis of price and volume capacity of a file server product, and file number is converted into processing time of introducing tired storage technology on the basis of data transfer rate and metadata writing speed. For the second case, cumulative file size is converted into accuracy of approximate calculation, and file number is converted into reduction of calculation time on the basis of time for selecting files and time for summing up file size of selected files. For the third case, cumulative file size is converted into cost effect of a file server product, and file number is converted into labor cost on the basis of work efficiency of manual confirmation in the case of deleting unnecessary files. Therefore our model contributes cost-effective file server from viewpoint of operational management. We believe that our model also can be utilized in many other simulation-based evaluations, such as access performance and file fragmentation [10, 17]. Furthermore, file access frequencies, fraction of duplicated content, or more other statistical characteristics can be modeled and be utilized for further efficiency of enterprise file servers.

References

1. Anderson E, Hall J, Hartline J, Hobbs M, Karlin AR, Saia J, Swaminathan R, Wilkes J (2001) An experimental study of data migration algorithms. In: Proceedings of the 5th international workshop on algorithm engineering, 145–158
2. Malhotra J, Sarode P, Kamble A (2012) A review of various techniques and approaches of data deduplication. Intr J Eng Pract 1(1):1–8
3. Agrawal N, Bolosky WJ, Douceur JR, Lorch JR (2007) A five-year study of file-system metadata. ACM Trans Storage 3(3):31–45
4. Downey AB (2001) The structural cause of file size distributions. In: Proceedings of the 2001 ACM SIGMETRICS international conference on measurement and modeling of computer systems
5. Evans KM, Kuenning GH (2002) A study of irregularities in file-size distributions. In: Proceedings of international symposium on performance evaluation of computer and telecommunication systems
6. Meyer DT, Bolosky WJ (2011) A study of practical deduplication. In: Proceedings of 9th USENIX conference on file and storage technologies
7. Satyanarayanan M (1981) A study of file sizes and functional lifetimes. In: Proceedings of the 8th ACM symposium on operating systems principles, 96–108
8. Barford P, Crovella M (1998) Generating representative web workloads for network and server performance evaluation. In: Proceedings of the 1998 ACM SIGMETRICS joint international conference on measurement and modeling of computer systems
9. Gibson T, Miller EL (1999) An improved long-term file-usage prediction algorithm. In: Proceedings of annual international conference on computer measurement and performance
10. SPEC SFS 2008 benchmark (2008); http://www.spec.org/sfs2008/
11. Matsumoto T, Onoyama T, Komoda N (2012) File size distribution model in enterprise file server toward efficient operational management. In: Proceedings of world congress on engineering and computer science 2012, 2, 1400–1404
12. Akaike H (1973) Information theory and an extension of the maximum likelihood principle. In: Proceedings of the 2nd international symposium on information theory, 267–281
13. Finney D (1941) On the distribution of a variable whose logarithm is normally distributed. J Roy Stat Soc 7:155–161
14. Sturges HA (1926) The choice of a class interval. J Am Stat Assoc 21(153):65–66
15. McLaren CE, Legler JM, Brittenham GM (1994) The generalized χ^2 goodness-of-fit test. The Statistician 43(2):247–258
16. Miller GR (1981) Simultaneous Statistical Inference, 2nd edn. Springer-Verlag, New York
17. Nakamura T, Komoda N (2009) Size adjusting pre-allocation methods to improve fragmentation and performance on simultaneous file creation by synchronous write. ISPJ Journal 50(11):2690–2698

Chapter 43
Dispersed Software Environment in Virtual Prototyping of Underground Mining Mechanical Systems

Jaroslaw Tokarczyk

Abstract The method of creation of computational models using the Multibody System (MBS) as well as the finite element method (FEM), is presented on the example of virtual prototyping in the dispersed software environment of the mechanical system used in underground coal mining industry. Arch yielding support together with highly loaded bearing systems for suspended monorails, generating dynamic loads during passage, is one of such systems. The results of dynamic loads to the support, tested in the laboratory according to the PN-92/G-15000/05 Standard, were compared with the selected criterial state, which includes support arches overload i.e. emergency braking of the transportation set, which carries large-size materials. The limit overloads causing local loss of arch support stability have been calculated.

Keywords FEM · MBS · Mining · Multibody system · Simulation · Virtual prototyping

43.1 Introduction

Bumping initiated by rock quake is the most dangerous natural hazard in Polish hard coal mines. Sudden reduce of roadway cross-section i.e. deformation of arch support can be the result. The load to arch support coming from rock mass is a subject of the research. There are the methods for simulation or calculation how the rock mass behave itself during a seismic quake [1] or cutting [2]. For that purpose the special software using Discrete Elements Method e.g., UDEC, 3DEC or FLAC is used. As regards the strength criterion [3], it is possible to generate

J. Tokarczyk (✉)
Institute of Mining Technology KOMAG, ul. Pszczyńska 37, 44-101 Gliwice, Poland
e-mail: jtokarczyk@komag.eu

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_43,

impact load in the special testing facility [4] due to the problems of identification of dynamic load of arch supports in in situ conditions. The requirements, which the testing facility has to meet as well as the method of loading the support, are given in standards [5].

In Poland, in all the mining underground plants the roadway yielding supports, which are made of steel arches connected with fasteners and clevises, are used. The support task is to keep stability of the roadway in a given time, what means keeping required dimensions of roadway cross-section as well as protection working people and machines against falling rock pieces and against roof falls. In Fig. 43.1 3D geometrical model of yielding arch support, created in CAD software environment, is given.

The arches of the support are pressed to each other by clevises generating static friction force between them. Possibility of suspension of additional equipment like hoisting winches or suspended monorail tracks, which cause dynamic load, is another function realized by the yielding arch supports. The Multi-Body System was used for simulation of load values.

In Poland, according to current regulations the amount of static load to single frames of arch support cannot exceed 40 kN [6]. Identification of dynamic overload, in a result of suspended monorail passage, acting on a single frame of arch roadway support, is one of simulation objectives. The monorails are used for transportation of materials and miners. They became dynamically developing technical mean used in the Polish underground mining industry [7]. However, increasing weight of transported large-size loads causes increase of overloads to suspended monorails tracks, what means higher load to rail joints, suspensions and finally to the frames of arch supports.

Identification of places of the highest strain, where damage can happen during operation, is the objective of virtual prototyping for the strength criterion at the stage of designing or verification of the mechanical system consisting of such components as a transportation unit, a railway route and arches of roadway support. Active forces and reaction forces during machine operation should not result in yielding of the material, which the machine components are made of. That is

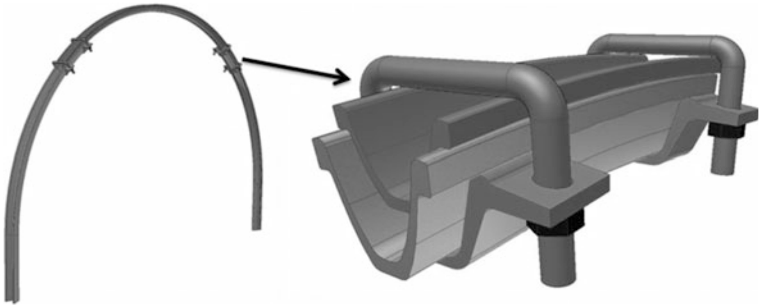


Fig. 43.1 Geometrical model of yielding arch support

why in calculation models linear models of materials are most often used. Nonlinearity of calculations is a result of contact phenomenon as well as due to complex load state of the computational model. General procedure of strength calculations is given in Fig. 43.2.

Designing process starts from a development of preliminary design, which is then verified using the special software tools. At this stage improvement of a design, except of elimination of design errors, is very important. For the selected criterial states the load conditions are identified. In the case of complex load conditions, additional initial calculations are made. In the case of dynamics problems or complex mechanical systems in the virtual prototyping process, the dispersed software environment is used. It also required the migration of computational models within this environment [8]. For instance, method for properly simulation of load flow between components of a mechanism is Multibody System (MBS). It enables determination of such amount as: force (*F*), force momentum (*M*), acceleration (*a*), speed (*v*). In the case of elastic-damping joints, it is necessary to determine stiffness coefficient (*c*) as well as damping coefficient (β), which will be used in the computational model. Then the results as input data are entered as boundary conditions for strength calculations, Fig. 43.3.

The obtained results of the strength calculation enable to optimize the design. Most often minimization of weight, at limitation of maximal reduced stresses and limitation of maximal displacements at selected points of the structure is the target function. Then current design is modified. In many cases the process has feedback loop character. After achievement of required parameters the mechanical system is designed, which is the base for development of design documentation.

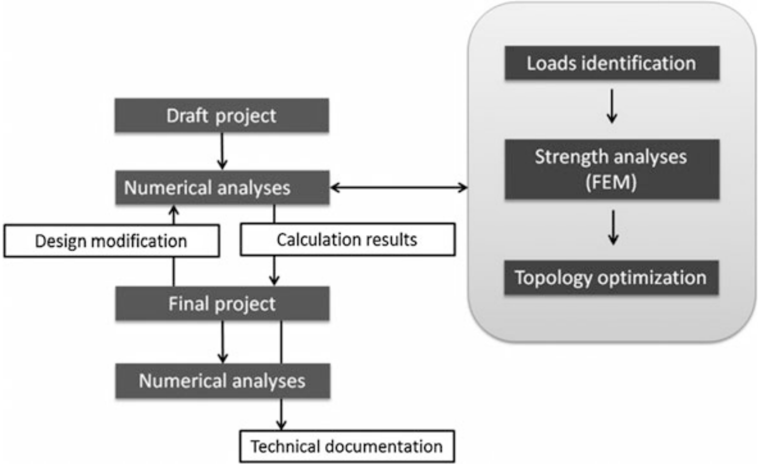


Fig. 43.2 Algorithm of operations during virtual prototyping of mechanical system

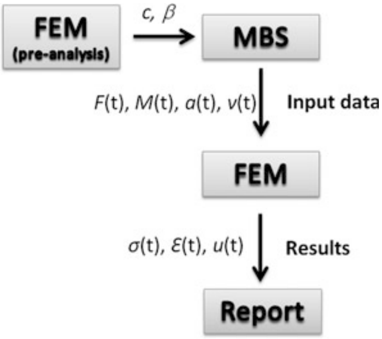


Fig. 43.3 Data flow between MBS and FEM software

43.2 Simulation of Emergency Braking of Suspended Transportation Set

The simulation was made using the MBS method. To determine action of high-loaded bearing units to arch support frames, the simulation of emergency braking of the transported load was made for the following parameters:

- Weight 15 t.
- Speed at the moment of starting braking 3 m/s.
- Static braking force: 60 kN.
- Inclination 10°.

Computational model for the railway route includes rails of the suspended monorail track connected with spherical and rotational joints, Fig. 43.4.

Totally the computational model consists of: 64 rigid bodies, 4 rotational joints, 44 spherical joints, 2 translational joints, 3 fixed joints, 22 elastic-dampening elements, 60 models of contacts. The model has 188 degrees of freedom (DOF), Fig. 43.5.

Then a single braking car with actuating components of the braking system was placed on the track, Fig. 43.6.

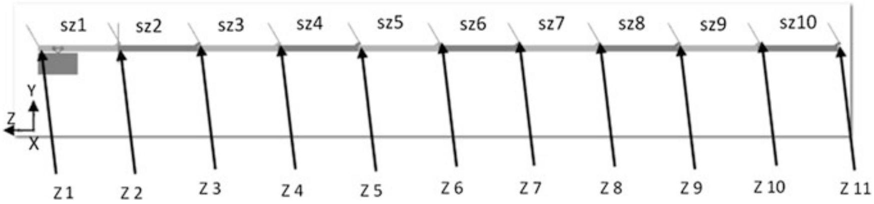


Fig. 43.4 MBS global model of suspended monorail track

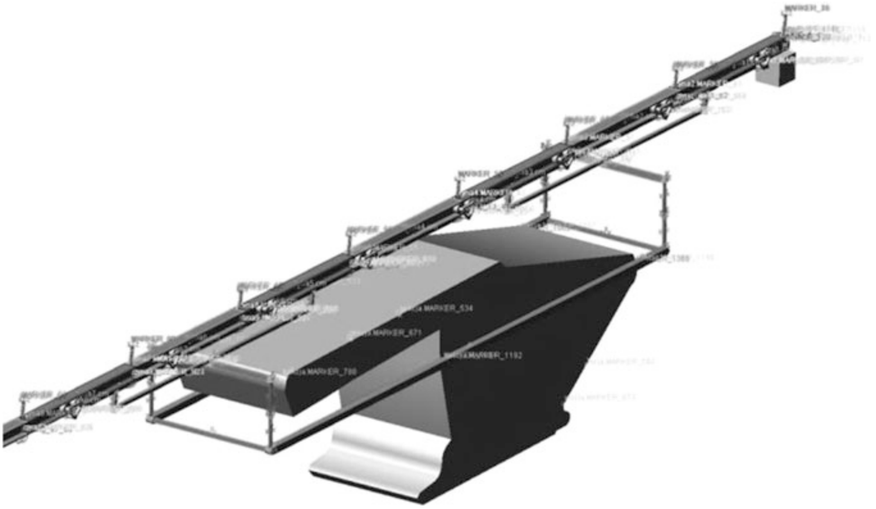


Fig. 43.5 MBS computational model of transportation set

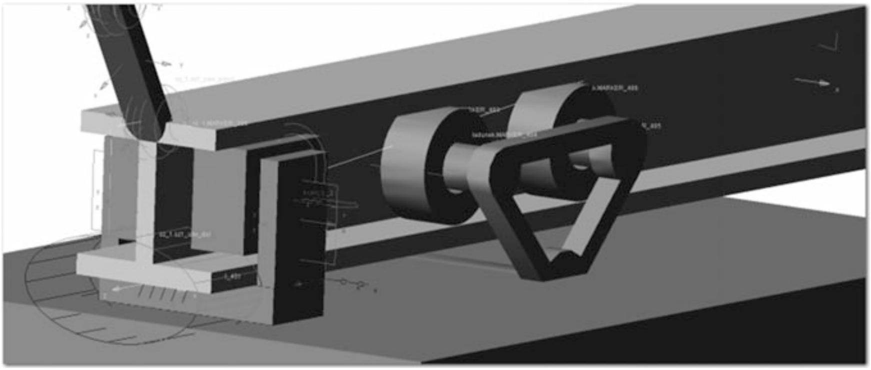


Fig. 43.6 Isometric view of transportation car with braking components

The simulation was carried out in two stages:

- Determination of so-called stable equilibrium.
- Acceleration of the bearing set and its braking.

After reaching a speed of 3 m/s, the force pressing brake blocks to the rail is released causing sudden braking of the transported load and in consequence overload to the suspensions of the track. In the result of rapid release of brakes in the braking car unstable dislocation of large-size load starts, Fig. 43.7.

On the basis of MBS calculations the forces in suspensions in function of time were identified. In Fig. 43.8 time process of forces in the suspension No. Z4 is given.

From the above diagram it results that maximal force in the suspension was 75 kN, what means 50 % of dynamic overload. Then the diagram of the force in

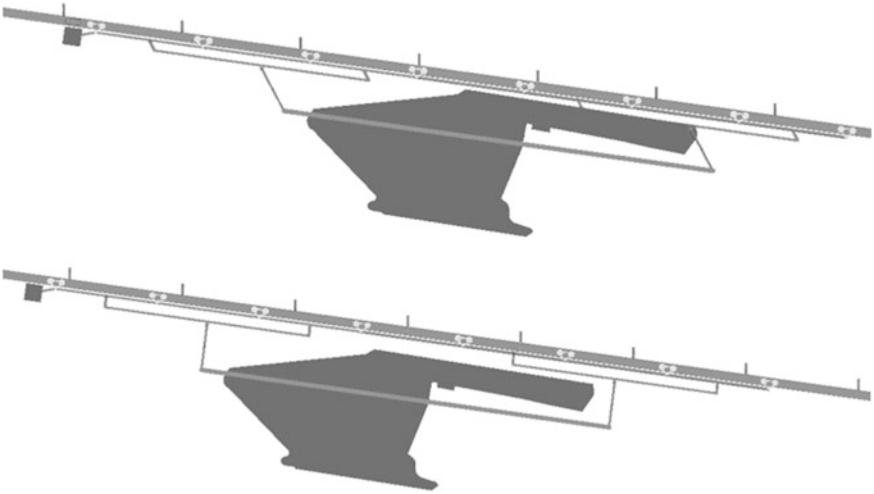


Fig. 43.7 Unstable behavior of load during emergency braking

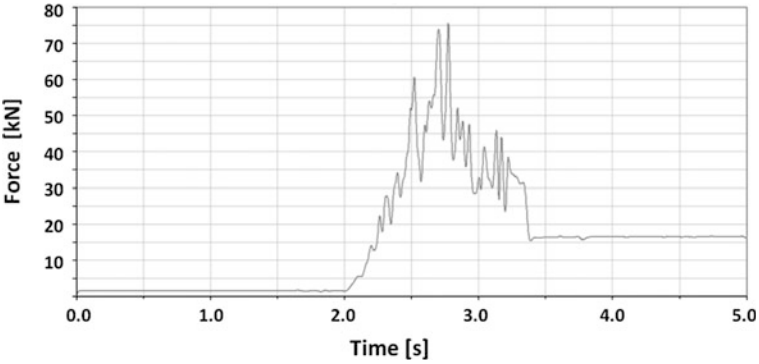


Fig. 43.8 Resultant force of Z4 suspension fixation

suspension was exported as input data to the FEM pre-processor environment and it was defined as so-called time dependent field.

43.3 Simulation of Impact Load to the Yielding Arch Support

Calculations objective was to compare dislocation of single arch of roadway support under dynamic load from the following sources:

- In the stand test for testing support frames [9].
- Emergency braking of the transportation set.

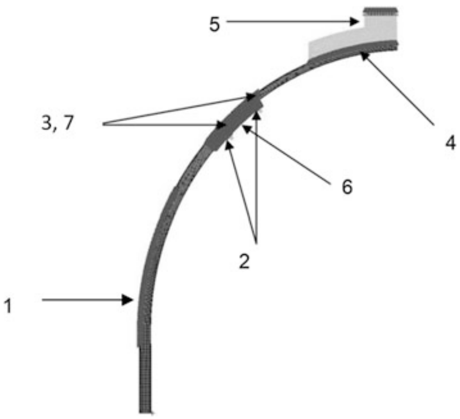


Fig. 43.9 Computational model of yielding arch support [9]

As the load is dynamic and of short duration, MSC.Patran/Dytran software, which enables explicit numerical analyses [10], was used in strength calculations. Possibility of arch support yielding, i.e. possibility of movement of side wall arch against roof arch, was introduced to the computational model. Such a movement is possible after exceeding frictional forces between the arches. For that purpose the models of contacts between side wall arch and roof arch were added. In the computational model the following models of contacts, presented in Fig. 43.9, were added:

- Side wall arch—side wall (1).
- Clevis—arch of the support (2).
- Rod element—arch of the support (3).
- Testing rig (beater)—arch of the support (4).
- Weight—testing rig (beater) (5).
- Side wall arch—roof arch (6).
- Rod element—clevis (7).

The results of total displacement of arch support frame have proved that impact load from falling mass in the testing facility had more severe effects (PHASE I) than load from emergency braking of transported mass (PHASE II). This can be seen in Fig. 43.10.

Maximal total displacement of the support after impact was about 40 mm, what resulted in yielding the frames against each other as well as plastic strains in roof arch.

Map of plastic strain of typical yielding arch support composed of canopy arch and wall arch is presented in Fig. 43.11.

Emergency braking caused support displacement of 2–3 mm. Only elastic strains occurred.

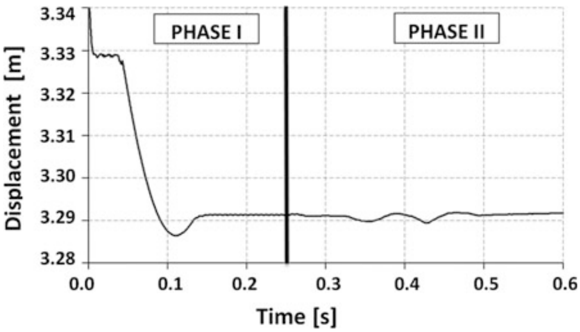


Fig. 43.10 Diagram of vertical displacement of the top part of the roof arch under dynamic load

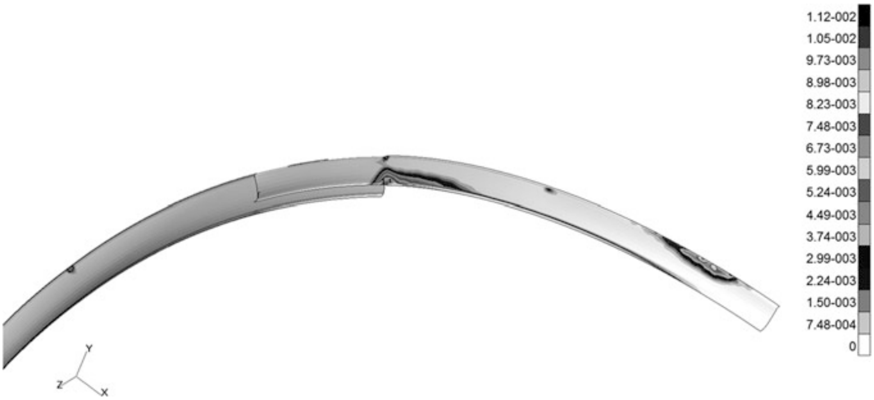


Fig. 43.11 Map of plastic strain of part of arch yielding support under dynamic load

43.4 Simulation of Damage of Arch Yielding Support Under Impact of Dynamic Load

Identification of dynamic critical load, from emergency braking of the transportation set, acting symmetrically on the frames of arch support, causing local loss of stability, i.e. arch buckling, was the simulation objective. Numerical calculations of explicit type were made to multiply (1, 2, 3 and 4) the load obtained from MBS analysis. The analysis was made using MSC.Patran/Dytran software.

In the computational model the following elastic-and-plastic material model with linear hardening was assumed:

- Density: 7850 kg/m^3.
- Elastic modulus: 205 GPa.
- Hardening modulus: 2.7 GPa.

- Poisson ratio: 0.3.
- Yield stress: 340 MPa.
- Max. plastic strain: 17 %.

In Fig. 43.12 diagrams of displacements of the node which is in the arch support axis for overloads 1,2 and 3 are given.

Overload 4, at which the maximal loading force reaches about 300 kN, causes local loss of stability of side wall arch (buckling) what leads to loss of load-bearing capacity of the support arches and collapse of the support, Fig. 43.13.

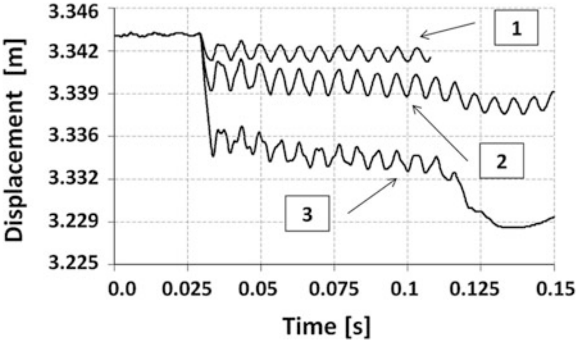


Fig. 43.12 Maximal displacements of arch of the support for impact overloads 1, 2 and 3

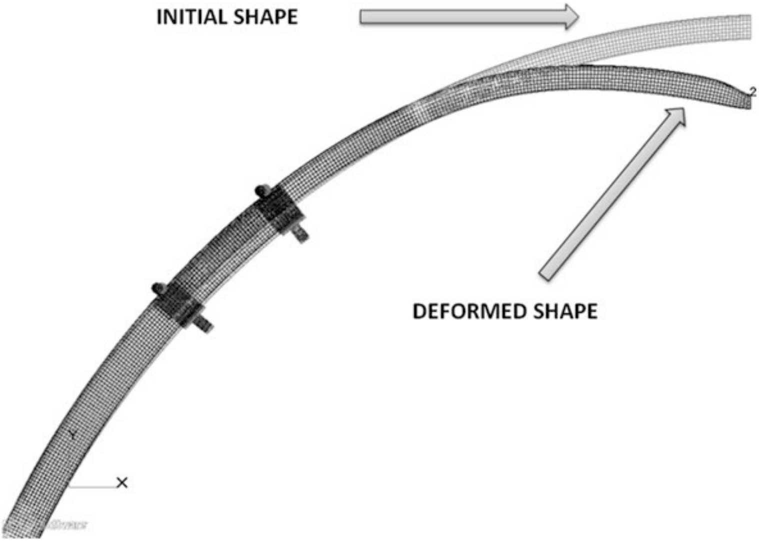


Fig. 43.13 Deformation of the arch yielding support for overload 4

43.5 Summary

Use of dispersed software environment for virtual prototyping of mechanical systems causes the necessity of data transfer between computer programs. It is also necessary to select dedicated software programs, which enable numerical simulations which will include the required physical phenomena. Additionally number of calculation tasks requires their parameterization and computer programs enabling at least partial automation of import/export operations between the calculation programs.

For the presented example the results clearly indicate that impact loads of energy about 30 kJ cause movement of the support frames against each other what results in reduction of the support height. The permissible load which, according to the regulations, can be applied to a single arch of the support i.e. 40 kN does not cause yielding the material of which the support is made. Overloading over 3.5 causes exceedance of yielding point and in consequence collapse of the support. Impact load can also lead to buckling of the support arches. State-of-the-art programs of class CAE enable defining any model of material and its verification in the investigated problem. They are broadly used in process of designing of new machines and equipment.

References

1. Kwaśniewski M (2003) Metody numerycznego modelowania rozprzestrzeniania energii sejsmicznej w górotworze. Praca naukowo—badawcza NB-146/RG-4/2003. Gliwice, 2003. (In Polish)
2. Labra C, Rojek J, Oñate E, Zarate F (2008) Advances in discrete element modelling of underground excavations. Acta Geotechnica 3(4):317–322
3. Winkler T, Tokarczyk J (2010) Multi-criteria assessment of virtual prototypes of mining machines. Proceedings: WCECS 2010, World congress on engineering and computer science, vol II, San Francisco, USA, 20–22 October, 2010 p 1149–1153
4. Pacześniowski K., Pytlik A., Radwańska E. (2007) Testing-stand tests of mine support elements with dynamic loads. Conference materials of the Polish Mining Congress 2007
5. PN-92/G-15000/05: Roadway support with susceptible timber frames made of special sections. Arches open frames set Tests. Polish standard. (In Polish)
6. Regulation of the Minister of Economy of the Republic of Poland: Rozporządzenie Ministra Gospodarki z dnia 28 czerwca 2002 r. w sprawie bezpieczeństwa i higieny pracy, prowadzenia ruchu oraz specjalistycznego zabezpieczenia przeciwpożarowego w podziemnych zakładach górniczych (Dz.U. 2002 nr 139 poz. 1169 wraz z późn. zm.). (In Polish)
7. MINTOS (2007) Contract no. RFCR-CT-2007-00003, “Improving Mining Transport Reliability”. European Project 2007–2010
8. Tokarczyk J (2012) Migration of computational models in virtual prototyping of complex mechanical systems. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 October, 2012, San Francisco, USA, p 1334–1337
9. Tokarczyk J, Turewicz K, Smolnik G, Rotkegel M (2010) (2010) Numerical analysis of impact load of arch yielding support. J KONES Powertrain Transp 17(1):455–464
10. MSC Software Corporation. www.mscsoftware.com. Last access 7.01.2013

Chapter 44
An Integrated Approach Based on 2-Tuple Fuzzy Representation and QFD for Supplier Selection

Mehtap Dursun and E. Ertugrul Karsak

Abstract With its need to trade-off multiple criteria exhibiting vagueness and imprecision, supplier selection is an important multi-criteria decision making problem. Vague and imprecise judgments inherent in numerous features of supplier selection justify the use of linguistic assessments rather than exact numerical values. In this chapter, a novel fuzzy multi-criteria group decision making approach integrating fusion of fuzzy information, 2-tuple linguistic representation model, and quality function deployment (QFD) is proposed for supplier selection. The developed fuzzy decision making approach employs ordered weighted averaging (OWA) operator and the aggregation process is based on combining information by means of fuzzy sets on a basic linguistic term set (BLTS). Afterwards, the collective performance values are transformed into linguistic 2-tuples to rectify the problem of loss of information encountered using other linguistic approaches. A supplier selection problem reported in an earlier study is used to illustrate the computational procedure of the proposed framework.

Keywords Decision support systems · Fusion of fuzzy information · OWA · QFD · Supplier selection · 2-tuple linguistic modeling

M. Dursun · E. E. Karsak (✉)
Industrial Engineering Department, Galatasaray University,
Ortakoy, 34357 Istanbul, Turkey
e-mail: ekarsak@gsu.edu.tr

M. Dursun
e-mail: mdursun@gsu.edu.tr

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_44,

44.1 Introduction

Supplier selection is considered as one of the key issues faced by operations and purchasing managers to sharpen the company's competitive advantage. As organizations become more dependent on their suppliers, the consequences of poor decisions on the selection of individual suppliers and the determination of order quantities to be placed with the selected suppliers become more severe [1]. Supplier selection decisions are complicated by the fact that various criteria must be considered in decision making process. Dickson [2] conducted one of the earliest works on supplier selection and identified 23 supplier attributes that managers consider when choosing a supplier.

Most of the research on supplier selection focuses on the quantifiable aspects of the supplier selection decision such as cost, quality, and delivery reliability. However, as firms become involved in strategic partnerships with their suppliers, a new set of supplier selection criteria, which are difficult to quantify, needs to be considered. Fuzzy set theory is an effective tool to address uncertainty in decision making. In the literature, there are several studies that use various fuzzy decision making techniques to evaluate suppliers.

A number of authors have used fuzzy mathematical programming approaches. Amid et al. [3] developed a fuzzy multi-objective model for supplier selection. Chen [4] developed a fuzzy mixed integer programming approach to account for multiple criteria and vagueness with the supplier selection decisions in the rebuy purchasing situation. More recently, Díaz-Madroñero et al. [5] addressed the supplier selection problem with fuzzy goals. A multi-objective model, which aims to minimize the total order costs, the number of rejected items and the number of late delivered items simultaneously, was developed.

Several studies have focused on the use of fuzzy multi-attribute decision making (MADM) techniques for supplier selection process. Wang et al. [6] proposed fuzzy hierarchical TOPSIS (technique for order preference by similarity to ideal solution) for supplier selection. Awasthi et al. [7] used fuzzy TOPSIS for evaluating environmental performance of suppliers. Sanayei et al. [8] proposed fuzzy multi-criteria optimization and compromise solution (VIKOR) method in order to select the suitable supplier in a supply chain system. Sevkli [9] developed fuzzy ELECTRE (ELimination Et Choix Traduisant la REalité) method for supplier selection, and compared the results obtained from crisp and fuzzy ELECTRE methods. Lately, Kang et al. [10] proposed fuzzy analytic network process (ANP) to solve the supplier selection problem. The developed model was implemented in an integrated circuit packaging company.

Some papers have proposed the use of 2-tuple fuzzy linguistic representation model. Wang [11] used 2-tuple fuzzy linguistic representation model to determine the overall supplier performance with dynamic supply behaviors. Wang [12] developed a model based on 2-tuple fuzzy linguistic representation model to evaluate the supplier performance.

Lately, few researchers have employed the quality function deployment (QFD) in supplier selection. Onesime et al. [13] proposed a supplier selection methodology based on QFD, analytic hierarchy process (AHP), and preemptive goal programming. AHP was used to measure the relative importance weights of supplier requirements and to assess the evaluation scores of candidate suppliers. Preemptive goal programming model was employed to assign order quantities to the suppliers. Bevilacqua et al. [14] constructed a house of quality to determine the features that the purchased product should possess to satisfy the customers' requirements. Afterwards, the potential suppliers were evaluated against the relevant supplier assessment criteria. Lately, Bhattacharya et al. [15] combined AHP with QFD to rank and then select candidate-suppliers under multiple, conflicting nature criteria environment.

Although previously reported studies developed approaches for supplier selection process, further studies are necessary to integrate imprecise information into the analysis, regarding the importance of purchased product features, relationship between purchased product features and supplier assessment criteria, and dependencies between supplier assessment criteria. With its need to trade-off multiple criteria exhibiting vagueness and imprecision, supplier selection is a highly important group decision making problem.

In this chapter, a fuzzy multi-criteria group decision making approach based on the concepts of fusion of fuzzy information, 2-tuple linguistic representation model, and QFD is proposed. The proposed method identifies how well each supplier characteristic accomplishes meeting the requirements established for the product being purchased by constructing a house of quality, which enables the relationships among the purchased product features and supplier assessment criteria to be considered. Moreover, this method enables the managers to deal with heterogeneous information, and thus, allows for the use of different semantic types by the decision-makers. The proposed decision making approach uses the ordered weighted averaging (OWA) operator to aggregate decision makers' opinions. The OWA operator is a common generalization of the three basic aggregation operators, i.e. max, min, and arithmetic mean.

The rest of the chapter is organized as follows. Sections 44.2 and 44.3 present the preliminaries concerning fusion of fuzzy information approach and 2-tuple fuzzy linguistic representation model, respectively. In Sect. 44.4, the fuzzy decision making framework is delineated. The application of the fuzzy decision making framework to supplier selection problem is expressed in Sect. 44.5. Finally, concluding remarks are given in Sect. 44.6.

44.2 Preliminaries

Fuzzy set theory, which was introduced by Zadeh [16] to deal with problems in which a source of vagueness is involved, has been utilized for incorporating imprecise data into the decision framework. A fuzzy set $\tilde{A}$ can be defined

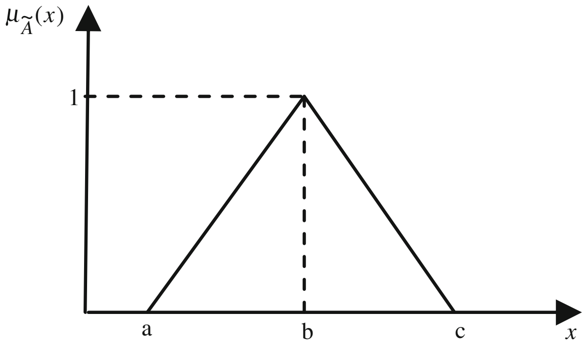


Fig. 44.1 A triangular fuzzy number $\tilde{A}$

mathematically by a membership function $\mu_{\tilde{A}}(x)$, which assigns each element x in the universe of discourse X a real number in the interval [0,1].

A triangular fuzzy number $\tilde{A}$ can be defined by a triplet (a, b, c) as illustrated in Fig. 44.1.

Triangular fuzzy numbers are appropriate for quantifying the vague information about most decision problems including supplier selection. The primary reason for using triangular fuzzy numbers can be stated as their intuitive and computational-efficient representation. A linguistic variable is defined as a variable whose values are not numbers, but words or sentences in natural or artificial language. The concept of a linguistic variable appears as a useful means for providing approximate characterization of phenomena that are too complex or ill defined to be described in conventional quantitative terms [17].

Fusion approach of fuzzy information, which was initially proposed by Herrera, Herrera-Viedma, and Martínez [18], is used to manage information assessed using both linguistic and numerical scales in a decision making problem with multiple information sources. This approach is carried out in two phases:

Phase 1. Making the information uniform: The performance values expressed using multi-granularity linguistic term sets are converted (under a transformation function) into a specific linguistic domain, which is a basic linguistic term set (BLTS), chosen so as not to impose useless precision to the original evaluations and to allow an appropriate discrimination of the initial performance values. The transformation function is defined as follows [18]:

Let $A = \{l_0, l_1, \ldots, l_p\}$ and $S_T = \{s_0, s_1, \ldots, s_g\}$ be two linguistic term sets, such that $g \geq p$. Then, the transformation function, τ_{AS_T}, is defined as

$$\begin{aligned} \tau_{AS_T} &= A \to F(S_T), \\ \tau_{AS_T}(l_k) &= \left\{ \left(s_i, \gamma_i^k\right) / i \in \{0, 1, \ldots, g\} \right\}, \forall l_k \in A, \\ \gamma_i^k &= \max_y \min\left\{ \mu_{l_k}(y), \mu_{s_i}(y) \right\}, \end{aligned} \tag{44.1}$$

where $F(S_T)$ is the set of fuzzy sets defined in S_T, and $\mu_{l_k}(y)$ and $\mu_{s_i}(y)$ are the membership functions of the fuzzy sets associated with the terms l_k and s_i, respectively.

Phase 2. Computing the collective performance values: For each alternative, a collective performance value is obtained by means of the aggregation of the aforementioned fuzzy sets on the BLTS that represents the individual performance values assigned to the alternative according to each information source [18]. Therefore, each collective performance value is a new fuzzy set defined on a BLTS. This chapter employs the OWA operator, initially proposed by Yager [19], as the aggregation operator. This operator provides an aggregation which lies in between the "and" requiring all the criteria to be satisfied, and the "or" requiring at least one of the criteria to be satisfied. Indeed, the OWA category of operators enables trivial adjustment of the ANDness and ORness degrees embedded in the aggregation [20]. The OWA operator differs from the classical weighted mean in that coefficients are not associated directly with a particular attribute but rather to an ordered position. It encompasses several operators since it can implement different aggregation rules by changing the order weights.

Let $A = \{a_1, a_2, \ldots, a_n\}$ be a set of values to be aggregated. The OWA operator F is defined as

$$F(a_1, a_2, \ldots, a_n) = \mathbf{w}\mathbf{b}^T = \sum_{j=1}^{n} w_j b_j, \tag{44.2}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ is a weighting vector, such that $w_i \in [0, 1]$ and $\sum_{i=1}^{n} w_i = 1$ and $\mathbf{b}$ is the associated ordered value vector, where $b_j \in \mathbf{b}$ is the jth largest value in A.

To apply the OWA operator for decision making, a crucial issue is to determine its weights. The weights of the OWA operator are calculated using fuzzy linguistic quantifiers, which for a non-decreasing relative quantifier Q, are given by

$$w_i = Q(i/n) - Q((i-1)/n),\ i = 1, \ldots, n. \tag{44.3}$$

The non-decreasing relative quantifier, Q, is defined as [18]

$$Q(y) = \begin{cases} 0 & , y < a \\ \frac{y-a}{b-a} & , a \le y \le b \\ 1 & , y > b \end{cases} \tag{44.4}$$

with $a, b, y \in [0, 1]$, and $Q(y)$ indicating the degree to which the proportion y is compatible with the meaning of the quantifier it represents. Some non-decreasing relative quantifiers are identified by terms 'most', 'at least half', and 'as many as possible', with parameters (a,b) given as (0.3,0.8), (0,0.5) and (0.5,1) respectively.

44.3 2-Tuple Fuzzy Linguistic Representation Model

The 2-tuple linguistic model that was presented by Herrera and Martínez [21] is based on the concept of symbolic translation. It is used for representing the linguistic assessment information by means of a 2-tuple that is composed of a linguistic term and a number. It can be denoted as (s_i, α) where s_i represents the linguistic label of the predefined linguistic term set S_T, and α is a numerical value representing the symbolic translation [22].

Let $r_1 = (s_c,\alpha_1)$ and $r_2 = (s_d,\alpha_2)$ be two linguistic variables represented by 2-tuples. The comparison of linguistic information represented by 2-tuples is carried out according to an ordinary lexicographic order as follows [23]:

- If $c < d$ then r_1 is smaller than r_2;
- If $c = d$ then
  - If $\alpha_1 = \alpha_2$ then r_1 and r_2 represent the same information;
  - If $\alpha_1 < \alpha_2$ then r_1 is smaller than r_2;
  - If $\alpha_1 > \alpha_2$ then r_1 is bigger than r_2.

In the following, we define a computational technique to operate with the 2-tuples without loss of information:

Definition 1 [24] Let $L = \left(\gamma_0, \gamma_1, \ldots, \gamma_g\right)$ be a fuzzy set defined in S_T. A transformation function χ that transforms L into a numerical value in the interval of granularity of S_T, $[0, g]$ is defined as

$$\chi : F(S_T) \rightarrow [0, g],$$

$$\chi(F(S_T)) = \chi\left(\left\{\left(s_j, \gamma_j\right), j = 0, 1, \ldots, g\right\}\right) = \frac{\sum_{j=0}^{g} j\gamma_j}{\sum_{j=0}^{g} \gamma_j} = \beta \tag{44.5}$$

where $F(S_T)$ is the set of fuzzy sets defined in S_T.

Definition 2 [21] Let $S = \left\{s_0, s_1, \ldots, s_g\right\}$ be a linguistic term set and $\beta \in [0, g]$ a value supporting the result of a symbolic aggregation operation, then the 2-tuple that expresses the equivalent information to β is obtained with the following function:

$$\Delta : [0, g] \rightarrow S \times [-0.5, 0.5),$$

$$\Delta(\beta) = \begin{cases} s_i, & i = \text{round}(\beta) \\ \alpha = \beta - i, & \alpha \in [-0.5, 0.5), \end{cases} \tag{44.6}$$

where ‘round’ is the usual round operation, s_i has the closest index label to ‘β’ and ‘α’ is the value of the symbolic translation.

Proposition 1 [21] *Let $S = \left\{s_0, s_1, \ldots, s_g\right\}$ be a linguistic term set and (s_i, α) be a 2-tuple. There is a Δ^{-1} function, such that, from a 2-tuple it returns its equivalent numerical value $\beta \in [0, g] \subset \Re$. This function is defined as*

$$\begin{aligned} \Delta^{-1} &: S \times [-0.5, 0.5) \rightarrow [0, g], \\ \Delta^{-1}(s_i, \alpha) &= i + \alpha = \beta. \end{aligned} \tag{44.7}$$

Definition 3 [12, 25] Let $x = \{(s_1, \alpha_1), \ldots, (s_n, \alpha_n)\}$ be a set of linguistic 2-tuples and $W = \{(w_1, \alpha_1^w), \ldots, (w_n, \alpha_n^w)\}$ be their linguistic 2-tuple associated weights. The 2-tuple linguistic weighted average $\bar{x}_l^w$ is calculated as

$$\bar{x}_l^w = \left(\left[(s_1, \alpha_1), (w_1, \alpha_1^w)\right] \ldots \left[(s_n, \alpha_n), (w_n, \alpha_n^w)\right]\right) = \Delta\left(\frac{\sum_{i=1}^{n} \beta_i \beta_{w_i}}{\sum_{i=1}^{n} \beta_{w_i}}\right) \tag{44.8}$$

with $\beta_i = \Delta^{-1}(s_i, \alpha_i)$ and $\beta_{w_i} = \Delta^{-1}(w_i, \alpha_i^w)$.

44.4 Fuzzy Decision Making Framework

In this section, a decision making approach that integrates the concepts of fusion of fuzzy information, 2-tuple linguistic representation model, and QFD is developed to address the supplier selection problem [26].

QFD aims at delivering value by focusing on prioritized customer needs, translating these into design requirements, and then communicating them throughout the organization in a way to assure that details can be quantified and controlled [27]. The reported benefits of QFD include better products or services that are highly focused and responsive to customer needs (CNs), developed in a shorter period of time with fewer resources. One shall also note the intangible benefits of QFD such as increased customer satisfaction, enhanced multi-disciplined teamwork, and structured basis for improved planning [28].

The basic concept of QFD is to translate the desires of customers into technical attributes (TAs), and subsequently into parts characteristics, process plans and production requirements. In order to establish these relationships, QFD usually requires four matrices each corresponding to a stage of the product development cycle. In this chapter, we focus on the first of the four matrices, also called the house of quality (HOQ).

The proposed approach considers the ambiguity resulting from imprecise statements in expressing relative importance of CNs, relationship scores between CNs and TAs, degree of dependencies among TAs, and the ratings of each potential supplier with respect to each TA by using fuzzy set theory. Moreover, utilization of the fusion of fuzzy information and the 2-tuple linguistic representation model enables decision-makers to deal with heterogeneous information, and rectify the problem of loss of information of other fuzzy linguistic approaches. The stepwise representation of the fuzzy MCDM framework is as follows [26]:

Step 1. Construct a decision-makers committee of Z $(z = 1, 2, \ldots, Z)$ experts. Identify the characteristics that the product being purchased must possess (CNs) in order to meet the company's needs and the criteria relevant to supplier assessment (TAs).

Step 2. Construct the decision matrices for each decision-maker that denote the fuzzy assessment to determine the CN-TA relationship scores, the relative importance of CNs, and the degree of dependencies among the TAs.

Step 3. Let the fuzzy value assigned as the relationship score between the lth CN $(l = 1, 2, \ldots, L)$ and kth TA $(k = 1, 2, \ldots, K)$, importance weight of the lth CN, and degree of dependence of the kth TA on the k'th TA for the zth decision-maker be $\tilde{x}_{klz} = (x^1_{klz}, x^2_{klz}, x^3_{klz}), \tilde{w}_{lz} = (w^1_{lz}, w^2_{lz}, w^3_{lz})$, and $\tilde{r}_{kk'z} = \left(r^1_{kk'z}, r^2_{kk'z}, r^3_{kk'z}\right)$, respectively. Convert $\tilde{x}_{klz}$, $\tilde{w}_{lz}$, and $\tilde{r}_{kk'z}$ into the basic linguistic scale S_T by using Eq. (44.1). The fuzzy assessment vector on S_T, the importance weight vector on S_T, and the degree of dependence vector on S_T, $F(\tilde{x}_{klz})$, $F(\tilde{w}_{lz})$, and $F(\tilde{r}_{kk'z})$ can be represented respectively as

$$F(\tilde{x}_{klz}) = (\gamma(\tilde{x}_{klz}, s_0), \gamma(\tilde{x}_{klz}, s_1), \ldots, \gamma(\tilde{x}_{klz}, s_6)), \ \forall\, k, l, z \tag{44.9}$$

$$F(\tilde{w}_{lz}) = (\gamma(\tilde{w}_{lz}, s_0), \gamma(\tilde{w}_{lz}, s_1), \ldots, \gamma(\tilde{w}_{lz}, s_6)), \ \forall\, l, z \tag{44.10}$$

$$F(\tilde{r}_{kk'z}) = (\gamma(\tilde{r}_{kk'z}, s_0), \gamma(\tilde{r}_{kk'z}, s_1), \ldots, \gamma(\tilde{r}_{kk'z}, s_6)), \ \forall\, k, k', z \tag{44.11}$$

In this study, the label set given in Table 44.1 is used as the BLTS.

Step 4. Aggregate $F(\tilde{x}_{klz})$, $F(\tilde{w}_{lz})$, and $F(\tilde{r}_{kk'z})$ to yield the fuzzy assessment vector $F(\tilde{x}_{kl})$, the importance weight vector $F(\tilde{w}_l)$, and the degree of dependence vector $F(\tilde{r}_{kk'})$. The aggregated parameters obtained from the assessment data of Z experts can be calculated respectively as

$$\tilde{x}_{kl}(s_m) = \phi_Q(\gamma(\tilde{x}_{kl1}, s_m), \gamma(\tilde{x}_{kl2}, s_m), \ldots, \gamma(\tilde{x}_{klz}, s_m)), \ \forall\, k, l, z \tag{44.12}$$

$$\tilde{w}_l(s_m) = \phi_Q(\gamma(\tilde{w}_{l1}, s_m), \gamma(\tilde{w}_{l2}, s_m), \ldots, \gamma(\tilde{w}_{lz}, s_m)), \ \forall\, l, m \tag{44.13}$$

$$\tilde{r}_{kk'}(s_m) = \phi_Q(\gamma(\tilde{r}_{kk'1}, s_m), \gamma(\tilde{r}_{kk'2}, s_m), \ldots, \gamma(\tilde{r}_{kk'z}, s_m)), \ \forall\, k, k', m \tag{44.14}$$

where ϕ_Q denotes the OWA operator whose weights are computed using the linguistic quantifier, Q. Then, the fuzzy assessment vector on S_T with respect to the

Table 44.1 Label set [29]

Label set	Fuzzy number
s_0:	(0,0,0.16)
s_1:	(0,0.16,0.33)
s_2:	(0.16,0.33,0.50)
s_3:	(0.33,0.50,0.66)
s_4:	(0.50,0.66,0.83)
s_5:	(0.66,0.83,1)
s_6:	(0.83,1,1)

lth CN, $F(\tilde{x}_{kl})$, the importance weight vector on S_T, $F(\tilde{w}_l)$, and the degree of dependence vector on S_T, $F(\tilde{r}_{kk'})$, is defined as follows:

$$F(\tilde{x}_{kl}) = (\gamma(\tilde{x}_{kl}, s_0), \gamma(\tilde{x}_{kl}, s_1), \ldots, \gamma(\tilde{x}_{kl}, s_6)),\ \forall\, k, l \tag{44.15}$$

$$F(\tilde{w}_l) = (\gamma(\tilde{w}_l, s_0), \gamma(\tilde{w}_l, s_1), \ldots, \gamma(\tilde{w}_l, s_6)),\ \forall\, l \tag{44.16}$$

$$F(\tilde{r}_{kk'}) = (\gamma(\tilde{r}_{kk'}, s_0), \gamma(\tilde{r}_{kk'}, s_1), \ldots, \gamma(\tilde{r}_{kk'}, s_6)),\ \forall\, k, k' \tag{44.17}$$

Step 5. Compute the β values of $F(\tilde{x}_{kl})$, $F(\tilde{w}_l)$, and $F(\tilde{r}_{kk'})$, and transform these values into a linguistic 2-tuple by using formulations (44.5) and (44.6), respectively.

Step 6. Compute the original relationship measure between the kth TA and the lth CN, $\tilde{X}^*_{kl}$. Let $D_{kk'}$ denote the degree of dependence of the kth TA on the k'th TA. Then, according to Fung et al. [30], the original relationship measure between the kth TA and the lth CN should be rewritten as

$$\tilde{X}^*_{kl} = \sum_{k'=1}^{K} D_{kk'} \tilde{x}_{k'l} \tag{44.18}$$

where $\tilde{X}^*_{kl}$ is the actual relationship measure after consideration of the inner dependence among TAs. Note that the correlation matrix $\mathbf{D}$ is symmetric. A technical attribute has the strongest dependence on itself, i.e. D_{kk} is assigned to be 1. If there is no dependence between the kth and the k'th TAs, then $D_{kk'} = 0$. Benefiting from Eq. (44.18), the original relationship measure is obtained by employing 2-tuple linguistic weighted average.

Step 7. Calculate the 2-tuple linguistic weighted average for each TA.

Step 8. Construct the decision matrices for each decision-maker that denote the ratings of each potential supplier with respect to each TA.

Step 9. Apply Steps 3–5 to the ratings of each supplier obtained at Step 8.

Step 10. Calculate the 2-tuple linguistic weighted average for each supplier. The associated weights are considered as the 2-tuple weighted average for each TA computed at Step 7.

Step 11. Rank the suppliers using the rules of comparison of 2-tuples given in Sect. 44.3.

44.5 Illustrative Supplier Selection Example

A supplier selection problem addressed in an earlier work by Bevilacqua et al. [14] is used to test the effectiveness of the proposed fuzzy MCDM framework. The problem is summarized as follows:

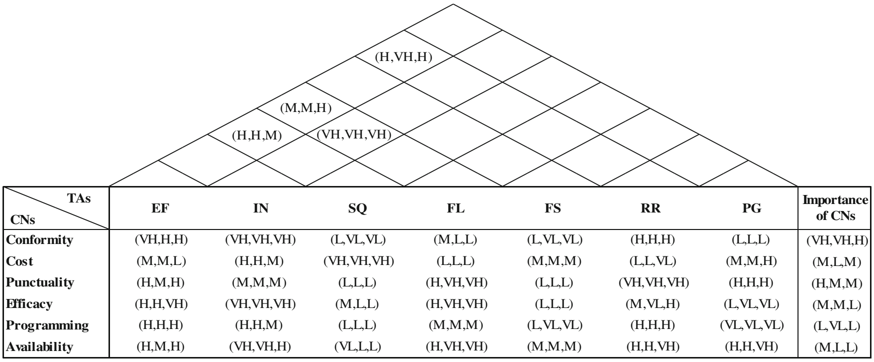
The analysis is performed for the selection of clutch plate suppliers for a medium-to-large enterprise that manufactures complete clutch coupling. There are ten suppliers who are in contact with the company. There are six fundamental characteristics (CNs) required of products or services purchased from outside suppliers by the company considered in this study. These can be listed as “product conformity”, “cost”, “punctuality of deliveries”, “efficacy of corrective action”, “programming of deliveries”, and “availability and customer support”. Seven criteria relevant to supplier assessment are identified as “experience of the sector (EF)”, “capacity for innovation to follow up the customer’s evolution in terms of changes in its strategy and market (IN)”, “quality system certification (SQ)”, “flexibility of response to the customer’s requests (FL)”, “financial stability (FS)”, “ability to manage orders on-line (RR)”, and “geographical position (PG)”. The evaluation is conducted by a committee of three decision-makers. The decision-makers used the linguistic variables given in Table 44.2 to denote the level of importance of each CN, the impact of each TA on CNs, and the ratings of the suppliers with respect to each TA.

The inner dependencies among TAs, which were ignored in the supplier selection problem addressed in Bevilacqua et al. [14], were considered in here as shown in Fig. 44.2.

First, the weights of CNs, the fuzzy assessment corresponding to the impact of each TA on each CN, and the inner dependencies among TAs are converted into the BLTS employing formulations (44.9)–(44.11). Next, by using the linguistic quantifier ‘most’ and the formulations (44.3) and (44.4), the OWA weights for three decision-makers are computed as $\mathbf{w} = (0.067, 0.666, 0.267)$. Then, the

Table 44.2 Linguistic term set

Very low (VL)	(0, 0.1, 0.2)
Low (L)	(0.2, 0.3, 0.4)
Medium (M)	(0.4, 0.5, 0.6)
High (H)	(0.6, 0.7, 0.8)
Very high (VH)	(0.8, 0.9, 1)



TAs CNs	EF	IN	SQ	FL	FS	RR	PG	Importance of CNs
Conformity	(VH,H,H)	(VH,VH,VH)	(L,VL,VL)	(M,L,L)	(L,VL,VL)	(H,H,H)	(L,L,L)	(VH,VH,H)
Cost	(M,M,L)	(H,H,M)	(VH,VH,VH)	(L,L,L)	(M,M,M)	(L,L,VL)	(M,M,H)	(M,L,M)
Punctuality	(H,M,H)	(M,M,M)	(L,L,L)	(H,VH,VH)	(L,L,L)	(VH,VH,VH)	(H,H,H)	(H,M,M)
Efficacy	(H,H,VH)	(VH,VH,VH)	(M,L,L)	(H,VH,VH)	(L,L,L)	(M,VL,H)	(L,VL,VL)	(M,M,L)
Programming	(H,H,H)	(H,H,M)	(L,L,L)	(M,M,M)	(L,VL,VL)	(H,H,H)	(VL,VL,VL)	(L,VL,L)
Availability	(H,M,H)	(VH,VH,H)	(VL,L,L)	(H,VH,VH)	(M,M,M)	(H,H,VH)	(H,H,VH)	(M,L,L)

Fig. 44.2 House of quality for the supplier selection problem

Table 44.3 Prioritization of the TAs using the proposed framework

CNs	Weights of CNs	TAs						
		EF	IN	SQ	FL	FS	RR	PG
Conformity	$(s_5,0.20)$	$(s_3,0.07)$	$(s_4,-0.23)$	$(s_2,0.24)$	$(s_4,-0.26)$	$(s_1,-0.16)$	$(s_4,0.24)$	$(s_2,-0.14)$
Cost	$(s_3,-0.12)$	$(s_3,-0.02)$	$(s_3,0.01)$	$(s_4,0.36)$	$(s_3,-0.12)$	$(s_3,0.01)$	$(s_2,0.21)$	$(s_3,0.01)$
Punctuality	$(s_3,0.14)$	$(s_4,0.07)$	$(s_4,0.04)$	$(s_3,-0.26)$	$(s_4,0.14)$	$(s_2,-0.14)$	$(s_5,-0.20)$	$(s_4,0.18)$
Efficacy	$(s_3,-0.12)$	$(s_4,-0.37)$	$(s_5,0.28)$	$(s_3,-0.06)$	$(s_5,0.07)$	$(s_2,-0.14)$	$(s_4,-0.42)$	$(s_1,-0.16)$
Programming	$(s_2,-0.28)$	$(s_3,0.44)$	$(s_4,-0.45)$	$(s_3,-0.21)$	$(s_4,-0.36)$	$(s_1,-0.16)$	$(s_4,0.18)$	$(s_1,-0.30)$
Availability	$(s_2,0.01)$	$(s_4,-0.22)$	$(s_5,0.20)$	$(s_3,-0.35)$	$(s_5,-0.05)$	$(s_3,0.01)$	$(s_4,0.20)$	$(s_4,0.32)$
2−tuple linguistic weighted average		$(s_3,0.44)$	$(s_4,0.08)$	$(s_3,-0.12)$	$(s_4,0.01)$	$(s_2,-0.22)$	$(s_4,-0.11)$	$(s_2,0.45)$

Table 44.4 2-Tuple linguistic ratings of suppliers

	EF	IN	SQ	FL	FS	RR	PG
Sup 1	$(s_3,-0.12)$	$(s_2,0.01)$	$(s_3,-0.12)$	$(s_3,0.14)$	$(s_3,0.01)$	$(s_2,-0.28)$	$(s_2,-0.16)$
Sup 2	$(s_4,0.18)$	$(s_3,0.14)$	$(s_3,0.14)$	$(s_5,0.20)$	$(s_5,0.35)$	$(s_1,-0.16)$	$(s_2,0.01)$
Sup 3	$(s_2,-0.16)$	$(s_5,0.20)$	$(s_5,0.20)$	$(s_2,-0.14)$	$(s_2,-0.14)$	$(s_2,-0.28)$	$(s_1,-0.16)$
Sup 4	$(s_3,-0.12)$	$(s_4,0.18)$	$(s_1,-0.16)$	$(s_1,-0.16)$	$(s_5,0.18)$	$(s_1,-0.30)$	$(s_4,0.32)$
Sup 5	$(s_5,0.35)$	$(s_5,0.35)$	$(s_1,-0.30)$	$(s_4,0.18)$	$(s_3,-0.12)$	$(s_2,0.01)$	$(s_1,-0.06)$
Sup 6	$(s_5,0.20)$	$(s_2,-0.28)$	$(s_3,-0.12)$	$(s_3,0.01)$	$(s_4,0.32)$	$(s_1,-0.30)$	$(s_1,-0.16)$
Sup 7	$(s_1,-0.16)$	$(s_3,0.01)$	$(s_5,0.35)$	$(s_5,0.20)$	$(s_2,-0.14)$	$(s_3,0.30)$	$(s_2,0.01)$
Sup 8	$(s_2,0.09)$	$(s_3,0.14)$	$(s_4,0.18)$	$(s_1,-0.16)$	$(s_4,0.18)$	$(s_5,0.35)$	$(s_1,-0.16)$
Sup 9	$(s_3,0.01)$	$(s_4,0.18)$	$(s_3,-0.12)$	$(s_2,0.09)$	$(s_1,-0.30)$	$(s_1,-0.16)$	$(s_1,-0.16)$
Sup 10	$(s_4,0.18)$	$(s_1,-0.16)$	$(s_3,0.14)$	$(s_2,-0.28)$	$(s_5,0.02)$	$(s_1,-0.06)$	$(s_{1,}-0.06)$

weights of CNs, the fuzzy assessment corresponding the impact of each TA on each CN, and the inner dependencies among TAs converted into the BLTS are aggregated employing formulations (44.12)–(44.17). The β values of these weights, ratings, and dependencies are computed and transformed into a linguistic 2-tuple using formulations (44.6) and (44.7), respectively. The original relationship measure between TAs and CNs is computed employing Eq. (44.18) and 2-tuple linguistic weighted average. Then, the 2-tuple linguistic weighted averages for each TA are calculated. The results are shown in Table 44.3.

The ratings of each supplier converted into the BLTS are aggregated and transformed into a linguistic 2-tuple as in Table 44.4. Finally, the 2-tuple linguistic weighted average for each supplier is calculated and the suppliers are ranked according to the 2-tuple linguistic weighted average score. The ranking order of the suppliers is obtained as Sup 2 $\succ$ Sup 5 $\succ$ Sup 7 $\succ$ Sup 8 $\succ$ Sup 3 $\succ$ Sup 6 $\succ$ Sup 1 $\succ$ Sup 4 $\succ$ Sup 9 $\succ$ Sup 10.

Table 44.5 summarizes the results obtained from the proposed decision algorithm. Supplier 2 is determined as the most suitable supplier, which is followed by supplier 5. While the fuzzy ranking principle of Bevilacqua et al. [14] cannot compare suppliers 1, 9, and 10, the methodology proposed in this study provides a

Table 44.5 Ranking of suppliers using the proposed algorithm

Suppliers	2-tuple linguistic weighted average score	Rank
Sup 1	$(s_2,0.46)$	7
Sup 2	$(s_3,0.32)$	1
Sup 3	$(s_3,-0.24)$	5
Sup 4	$(s_2,0.37)$	8
Sup 5	$(s_3,0.29)$	2
Sup 6	$(s_3,-0.44)$	6
Sup 7	$(s_3,0.22)$	3
Sup 8	$(s_3,-0.08)$	4
Sup 9	$(s_2,0.25)$	9
Sup 10	$(s_2,0.16)$	10

complete ranking of all suppliers. This is due to the minimization of the loss of information by using the 2-tuple fuzzy linguistic representation model.

44.6 Conclusion and Future Work

In this chapter, a fuzzy multi-criteria group decision making algorithm based on the concepts of fusion of fuzzy information, 2-tuple linguistic representation model, and QFD is presented to rectify the problems encountered when using classical decision making methods in supplier selection. The proposed decision framework enables both relationship between purchased product features and supplier assessment criteria, and inner dependencies between supplier assessment criteria to be taken into consideration. The decision making approach presented in this chapter disregards the troublesome fuzzy number ranking process, which may yield inconsistent results for different ranking methods, and as a result improves the quality of decision. Moreover, the decision methodology enables managers to deal with heterogeneous information, and thus, allows for the use of different semantic types by decision-makers.

Future research will address the implementation of the proposed methodology in real-world group decision making settings across diverse disciplines that can be represented in HOQ structure.

Acknowledgments This work was financially supported by Galatasaray University Research Fund under Grant 10.402.018.

References

1. Boer L, Labro E, Morlacchi P (2001) A review of methods supporting supplier selection. Eur J Purch Supply Manag 7:75–89
2. Dickson G (1966) An analysis of vendor selection systems and decisions. J Purch 2:28–41
3. Amid A, Ghodsypour SH, O'Brien C (2009) A weighted additive fuzzy multiobjective model for the supplier selection problem under price breaks in a supply Chain. Int J Prod Econ 121:323–332
4. Chen CM (2009) A fuzzy-based decision-support model for rebuy procurement. Int J Prod Econ 122:714–724
5. Díaz-Madroñero M, Peidro D, Vasant P (2010) Vendor selection problem by using an interactive fuzzy multi-objective approach with modified S-curve membership functions. Comput Math Appl 60(4):1038–1048
6. Wang JW, Cheng CH, Kun-Cheng H (2009) Fuzzy hierarchical TOPSIS for supplier selection. Applied Soft Computing 9:377–386
7. Awasthi A, Chauhan SS, Goyal SK (2010) A fuzzy multicriteria approach for evaluating environmental performance of suppliers. Int J Prod Econ 126:370–378
8. Sanayei A, Mousavi SF, Yazdankhah A (2010) Group decision making process for supplier selection with VIKOR under fuzzy environment. Expert Syst Appl 37:24–30

9. Sevkli M (2010) An application of the fuzzy ELECTRE method for supplier selection. Int J Prod Res 48(12):3393–3405
10. Kang HY, Lee AHI, Yang CY (2012) A fuzzy ANP model for supplier selection as applied to IC packaging. J Intell Manuf 23(5):1477–1488
11. Wang SY (2008) Applying 2-tuple multi-granularity linguistic variables to determine the supply performance in dynamic environment based on product-oriented strategy. IEEE Trans Fuzzy Syst 16:29–39
12. Wang WP (2010) A fuzzy linguistic computing approach to supplier evaluation. Appl Math Model 34:3130–3141
13. Onesime OCT, Xiaofei X, Dechen XZ (2004) A decision support system for supplier selection process. Int J Inf Technol Decis Mak 3(3):453–470
14. Bevilacqua M, Ciarapica FE, Giacchetta G (2006) A fuzzy-QFD approach to supplier selection. J Purchas Supply Manag 12:14–27
15. Bhattacharya A, Geraghty J, Young P (2010) Supplier selection paradigm: An integrated hierarchical QFD methodology under multiple-criteria environment. Appl Soft Comput 10:1013–1027
16. Zadeh LA (1965) Fuzzy sets. Inf Control 8(3):338–353
17. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning-I. Inf Sci 8(3):199–249
18. Herrera F, Herrera-Viedma E, Martínez L (2000) A fusion approach for managing multi-granularity linguistic term sets in decision making. Fuzzy Sets Syst 114:43–58
19. Yager RR (1988) On ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE T Syst Man Cyb 18:183–190
20. Zarghami M, Szidarovszky F (2008) Fuzzy quantifiers in sensitivity analysis of OWA operator. Comput Ind Eng 54:1006–1018
21. Herrera F, Martínez L (2000) A 2-tuple fuzzy linguistic representation model for computing with words. IEEE Trans Fuzzy Syst 8:746–752
22. Fan ZP, Feng B, Sun YH, Ou W (2009) Evaluating knowledge management capability of organizations: a fuzzy linguistic method. Expert Systems Appl 36:3346–3354
23. Herrera F, Martínez L (2001) A model based on linguistic 2-tuples for dealing with multigranular hierarchical linguistic contexts in multi-expert decision-making. IEEE Trans Syst, Man, Cybern—Part B: Cybern 31:227–234
24. Herrera F, Martínez L (2000) An approach for combining linguistic and numerical information based on 2-tuple fuzzy representation model in decision-making. Int J Uncertainty Fuzziness Knowl-Based Syst 8:539–562
25. Herrera-Viedma E, Herrera F, Martínez L, Herrera JC, López AG (2004) Incorporating filtering techniques in a fuzzy linguistic multi-agent model for information gathering on the web. Fuzzy Sets Syst 148:61–83
26. Dursun M, Karsak EE (2012) Supplier selection using an integrated decision making approach based on QFD and 2-tuple fuzzy representation. Lecture notes in engineering and computer science: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24-26 October, 2012, USA, p 1309–1315
27. Karsak EE (2008) Robot selection using an integrated approach based on quality function deployment and fuzzy regression. Int J Prod Res 46:723–738
28. Revelle JB, Moran JW, Cox CA (1998) The QFD Handbook. Wiley, New York
29. Jiang YP, Fan ZP, Ma J (2008) A method for group decision making with multi granularity linguistic assessment information. Inf Sci 178:1098–1109
30. Fung RYK, Tang J, Tu Y, Wang D (2002) Product design resources optimization using a non-linear fuzzy quality function deployment model. Int J Prod Res 40:585–599

Chapter 45
Development of a Novel Approach for Electricity Forecasting

Mehdi K. Moghaddam and Parisa A. Bahri

Abstract In this chapter an innovative method for one and seven-day forecast of electricity load is proposed. The new approach has been tested on three different cases from south-west Western Australia's interconnected system. They have been tested under the most realistic conditions by considering only minimum and maximum forecasts of temperature and relative humidity as available future inputs. Two different nonlinear approaches of neural networks and decision trees have been applied to fit proper models. A modified version of mean absolute percentage error (MMAPE) of each model over the test year is presented. By applying a developed criterion to recognize the dominant component of the electricity load, user of this work will be able to choose the most efficient forecasting method.

Keywords Commercial load · Decision trees · Industrial load · Load forecasting · Neural networks · Residential load · Signal reconstruction

45.1 Introduction

The complexities of nowadays electricity markets are enormous [1]. Electricity is traded based on bilateral contracts between energy providers and energy consumers. The role of electricity demand forecasting is very significant in present electricity markets and an accurate load forecast introduces significant savings in

M. K. Moghaddam · P. A. Bahri (✉)
School of Engineering and Information Technology, Murdoch University, Perth, WA 6150, Australia
e-mail: P.Bahri@murdoch.edu.au

M. K. Moghaddam
e-mail: Mehdi.kardehi@gmail.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_45,

costs and improvement in network reliability. Forecasts ranging from several hours to seven days are known as short term load forecasts (STLF). Applications of STLF are in dispatching and commitment of generators, load shedding and determining the market prices. Because of its importance many methods have been developed to perform STLF. A review of the previous methods are addressed in Refs. [2–4]. The focus of this chapter is on two methods of artificial neural networks (ANN), and decision trees. References [5, 6] have applied neural networks for hourly load forecasts. Applications of artificial neural networks in one-day load forecasts have been addressed in Refs. [7, 8]. Weather inputs have not been considered in Ref. [7], and Munkhjargal and Manusov [8] has assumed measured values of temperatures during the test period. Using the measured values of weather variables in load forecasting instead of simulated or reconstructed values, makes the method incapable of performing the forecast in real world problems when measured values of weather variables are not available at the time of forecasting. An adaptive neural network approach has been used for seven-day forecast of electricity load in Ref. [9], in which the weather variables have not been directly considered as inputs. Electricity load is decoupled into three different ranges of frequencies and each range is forecasted by one neural network. Deviation of the achieved load forecast from the real one defines the temperature sensitive component which is forecasted by another neural network. Unfortunately the method has been tested during the winter time in the north east of USA where, because of heavy snow and freezing cold weather during the winter time, electric heating is not common. According to Weron [1] temperature-load curve of such a kind of load makes a hockey stick shape with very small correlation between temperature and load in the cold season. Reference [10] applied decision trees to forecast the demand in retail sale. Decision trees have been introduced as a potential method to predict a one-day load in Spanish power systems in Ref. [11].

Although numerous methods have been proposed for the short-term forecasting of electricity load, there is no superior forecasting approach that can be applied on all the different systems [1]. The main reasons for that are the unique characteristics of each system, and also the different consumer behaviors. Such characteristics become more significant in spatial load forecasting applications. More vital information from the grid can be extracted by having the spatial load forecast in hand. As mentioned by Willis [12] a spatial forecast that covers all the regions of the service area can assure the planner that nothing has been missed in the utility transmission and distribution planning. Spatial load forecasting divides the service area into different regions and each region load has its own characteristics. In this work different load behaviors of a sample service area will be thoroughly investigated and the efficiency of two forecasting methods will be tested. By defining a load type determination criterion the best approach for the presented benchmarks will be selected. The content of this chapter is originated from the work prepared by the authors for the world congress on engineering and computer science 2012 [13].

45.2 Case Study

In real world problems the electricity load data are composed of residential, industrial and commercial components. To study the characteristics of each component individually, three different benchmarks will be used. A pure residential area, an industrial area, and a dominantly commercial area in south–west Western Australia have been selected as the case studies. The main reason for choosing a pure residential load for model testing is that there are no high autocorrelations of industrial load in the data, so the model should be able to capture all the behaviours of households. The selected residential area load is highly temperature sensitive. The region is called East Perth metropolitan area and consists of one 6.6 kV and six 22 kV distribution substations and the total of seventeen transformers. Kalgoorlie[1] area has been selected as the industrial load sample. Industrial loads have their own complexities; for example the random behaviour of a large industrial customer can dramatically deviate the real load from the forecasted one. The region consists of one 11 kV and three 33 kV distribution substations and the total of seventeen transformers. And finally, Perth CBD with nine substations and the total of twenty five transformers has been selected as the dominantly commercial load. Compared to residential loads, commercial loads exhibit higher temperature sensitivity in hot seasons and lower temperature sensitivity in cold seasons.

The load information of each individual transformer has been extracted from the database of Western Power, the company that is responsible for building, maintaining and operating SWIS[2] electricity grid. The load data are then added up to find the electricity load of the test regions. The weather data are provided by the Australian Bureau of Meteorology (BOM). The specifications of raw data for the short term forecasting framework are presented in Table 45.1.

45.3 Data Preparation

45.3.1 Removing Outliers and Missing Data Points

The raw input data are composed of seven-day minimum and maximum temperature and relative humidity forecasts and also the historical data of load, temperature and relative humidity. A plot of Kalgoorlie load data for fifteen years of observation is presented in Fig. 45.1. The presented figure contains the raw data. It can be seen that during the first 50,000 samples the load has been dramatically increased. The main reason for that is commissioning new industrial projects. In the future steps this part

[1] Also known as country goldfields.

[2] SWIS: South west interconnected system.

Table 45.1 Raw data specification

Data	Unit	Resolution	Start-date	End-date
Load	MW	Half an hourly	01-Apr-1995	01-Jan-2011
Temperature	°C	Hourly	01-Apr-1995	01-Jan-2011
Relative humidity	%	Hourly	01-Apr-1995	01-Jan-2011

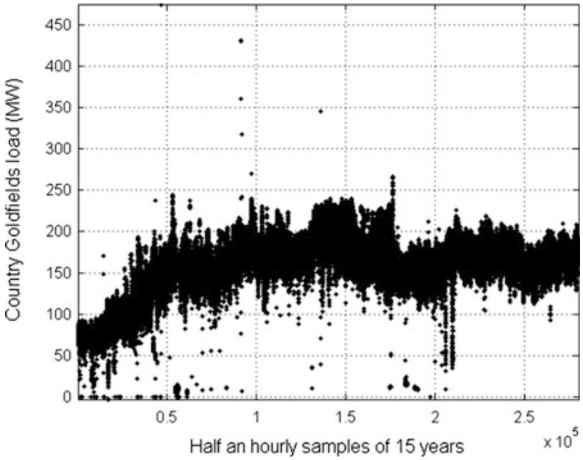


Fig. 45.1 Kalgoorlie raw data for fifteen years of observation

will be excluded from the industrial load data to increase the forecasting accuracy. As it can be seen, the raw data have some outliers. There also exist some missing data points in this figure that are not observable unless you zoom in.

After resolution adjustment, outliers and missing data points need to be removed. According to the available literature on electricity load consumption behaviours [14–16] and also the practical methods that are being used in electricity industries a transformer load in a specific time of a day is closely related to previous and future weeks[3] load data at the same time of the same day. Missing data points have been replaced by the average of previous and future weeks'load of the same hour of the same weekday.

As being recommended by Weron [1], one dimensional median filtering can be used for outliers removal. But median filtering by itself is not capable of removing all the outliers automatically. To capture normal outliers a short window should be used. A long window needs to be applied after that to capture outliers in a row. Human supervision is also required for outlier removals. The human expert can change the window sizes and investigate the data graphs and Q–Q[4] plots to make

[3] The correlation between the load data decreases as the number of weeks increases. In this study the maximum number of two future weeks and two previous weeks has been used for missing data estimation.

[4] Quantile–Quantile or Q–Q plot is a graph that shows the probability of two distributions against each other. By using Q–Q plots similarities and differences of two different distributions can be investigated.

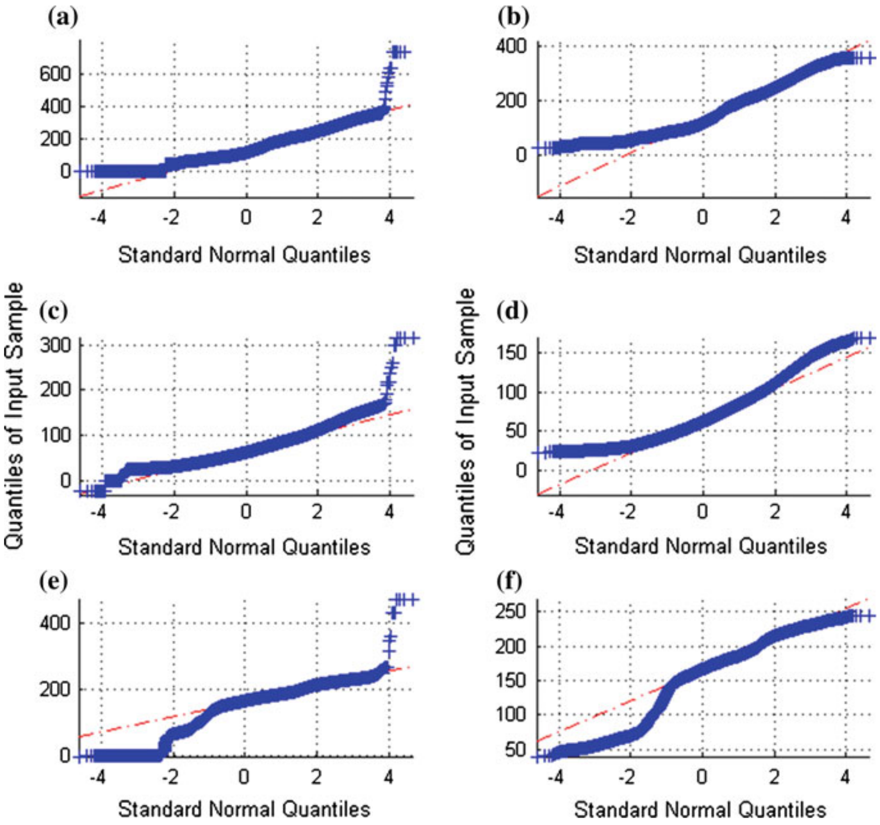


Fig. 45.2 Q–Q Plot of Load Data Versus Standard Normal, **a** Raw commercial load, **b** Commercial load without outliers and missing data points, **c** Raw residential load, **d** Residential load without outliers and missing data points, **e** Raw industrial load, **f** Industrial load without outliers and missing data points

sure that the outliers have been removed properly. Supervised one dimensional median filtering has been used for outliers' removal. Q–Q plots of the three systems, before, and after the outliers' removal step are presented in Fig. 45.2. Outliers can be easily identified in panel (a), (c) and (e) which contain the raw data of Perth CBD, East Perth and Kalgoorlie respectively. Panels (b), (d) and (f) confirm the capability of this method in outliers' removal.

45.3.2 Clustering and Signal Reconstruction

Clustering has been intensively used in load forecasting applications [17–19]. Because of the seasonal nature of weather variables and electricity consumption, clustering can help in deriving very important information out of the data set.

Weather forecasters are usually capable of giving out seven-day forecasts of weather variables in a limited resolution[5] of time. Because of the unpredictable nature of influential variables on weather systems, forecasts of beyond this horizon cannot be accurate enough to rely on.

The most important elements of weather for electricity demand forecasting are temperature and relative humidity [12]. In load forecasting applications the input data to the framework should be realistic and available at the time of running the framework. If not, the framework would become useless for practical applications. To avoid this issue, only the minimum and maximum values of temperature and relative humidity for seven days are considered as the future weather inputs of this framework.

To extract the weather distribution data out of the available minimum and maximum forecasts, historical data of temperature and relative humidity are clustered. Data sets of each cluster follow a similar pattern. By recognizing such a kind of pattern in the clusters and using the maximum and minimum forecasts of temperature and relative humidity, their signals are reconstructed. Figure 45.3 shows daily temperature of a representative cluster with a regular pattern to be easily seen in that. With the help of clustering, and using the available seven-day forecast of maximum and minimum temperature and relative humidity, weather signals can be accurately reconstructed as a feed to training models.

45.4 Behaviors of Residential, Industrial and Commercial Loads

Behaviors of residential, industrial and commercial loads are different. Although in the practical case the load can be a combination of all the three types, it is worthwhile to study the properties of each separately. In this section the ways that these loads can be distinguished from each other will be presented. A criterion will be proposed to recognize the dominancy of any of the mentioned types in a load data set. This criterion can then be used in places where the dominant type of load is not known.

45.4.1 Temperature Sensitivity

The following are the region's load versus temperature for 15 years of data. The white line in each graph roughly shows the regression between load and temperature

[5] Although sometimes these forecasts are available for every three hour of the following week, to avoid the loss of generality only minimum and maximum values of temperature and humidity are considered to be available to this framework at the time of forecasting.

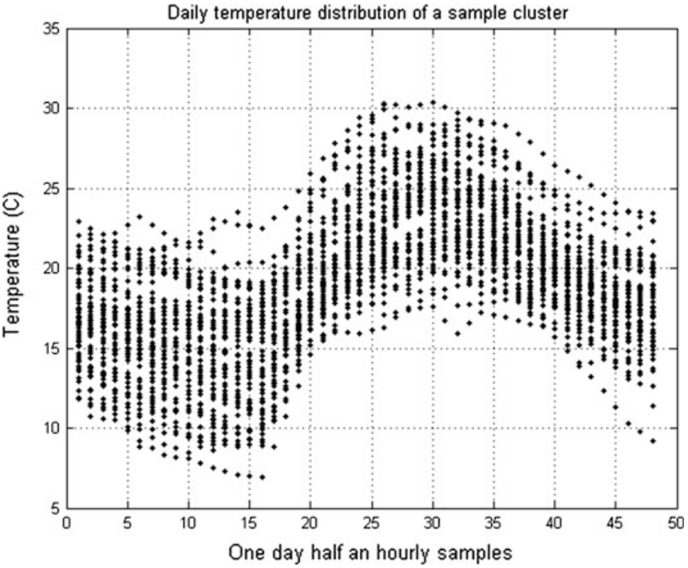
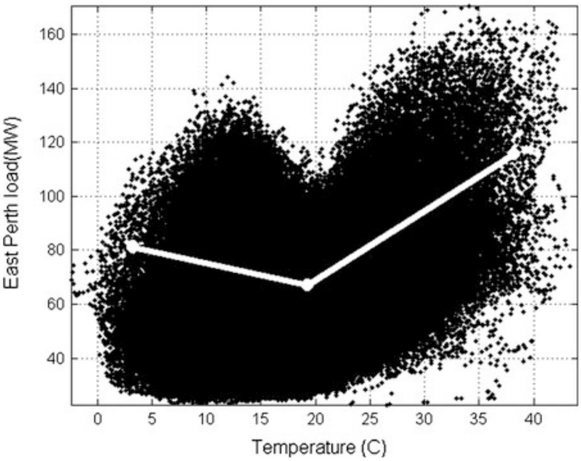


Fig. 45.3 Daily temperature distribution of a representative cluster from east perth

during the hot and cold seasons. In most of the cases, the slope of the line shows positive regression in the hot season and negative regression for the cold season.

Figure 45.4 shows the scatter plot of East Perth electricity consumption versus temperature over fifteen years of observation. It can be seen that, in the residential case, electricity consumption is very sensitive to temperature changes, and household cooling and heating demands strongly affect the load. The temperature correlation with load exhibits seasonal changes. At around 20 °C, which is known as the comfort region, the temperature-load correlation is close to zero. Positive

Fig. 45.4 Half an hourly electrical load consumption (MW) of the residential region versus temperature data (degrees celsius) from fifteen years of observation



regression for the hot season and negative regression for the cold season is very clear in this figure. The reason behind this type of graph is the fact that East Perth is mainly a residential area and people of that area use electricity both for cooling and heating purposes.

Figure 45.5 shows the scatter plot of the Perth CBD load versus temperature. The white lines in this plot are very similar to East Perth. There are two main differences between commercial loads and residential loads. Commercial loads drop dramatically after the business hours, and they usually have less heating demand and larger cooling demand. This fact increases the white line slope in the hot season and decreases the slope in the cold season. The main reason for this observation is the heating load which is generated by electronic devices inside commercial buildings.

Figure 45.6 shows the electricity load of country goldfields versus temperature. White lines are almost flat in this graph. That illustrates a negligible temperature sensitivity of load for this case. Irrespective of the outside temperature, load varies based on the factory demand. This confirms the fact that this region is dominated by industrial loads.

It can be concluded that more integration of residential load into a grid will introduce more temperature sensitivity and on the other hand more integration of industrial load will reduce it. The behavior of commercial loads is very similar to residential loads in the hot season and resembles industrial loads in cold seasons. It is very important to mention that such kind of conclusions can be only valid for the places where people use electricity for both cooling and heating purposes. If the users do not use electricity for heating purposes then the temperature-load regression during the winter will be zero (flat line) and the similar case will happen if the users do not use electricity for cooling purposes. The latter case is very rare but the example would be the customers who are using absorption chillers.

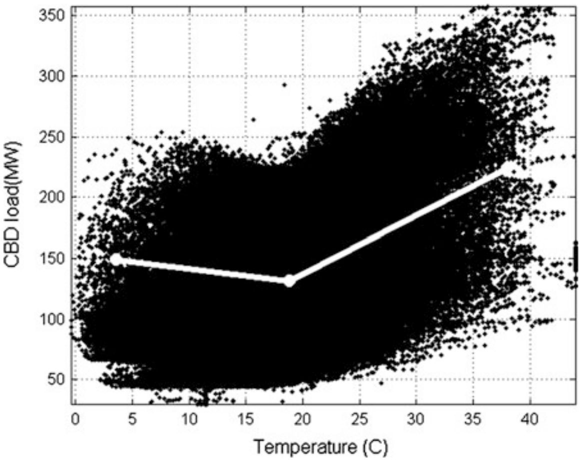


Fig. 45.5 Half an hourly electrical load consumption (MW) of the commercial region versus temperature data (degrees celsius) from fifteen years of observation

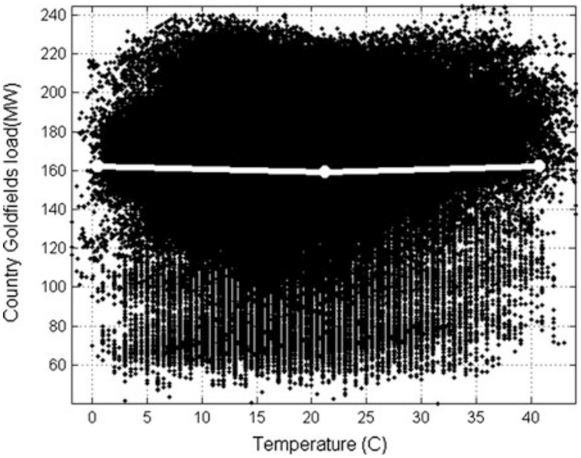


Fig. 45.6 Half an hourly electrical load consumption (MW) of the industrial region versus temperature data (degrees celsius) from fifteen years of observation

45.4.2 Data Distributions

Another way of distinguishing among the three types of loads is distribution analysis. Q–Q plots of the electricity load versus three different distributions are presented bellow.

Figures 45.7 and 45.8 illustrate the Q–Q plots of all three types of load versus Rayleigh (R) and generalized Pareto (GP) distributions. In both figures the best fit is for commercial load. It is not completely fitted for the residential load but it is a fairly good fit compared to the industrial one which shows a totally different distribution.

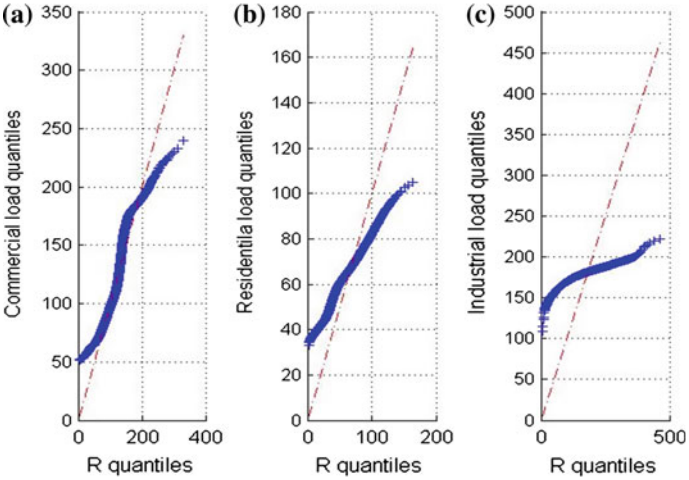


Fig. 45.7 **a** Q–Q plot of commercial load versus rayleigh distribution, **b** Q–Q plot of residential load versus rayleigh distribution, **c**. Q–Q plot of industrial load versus rayleigh distribution

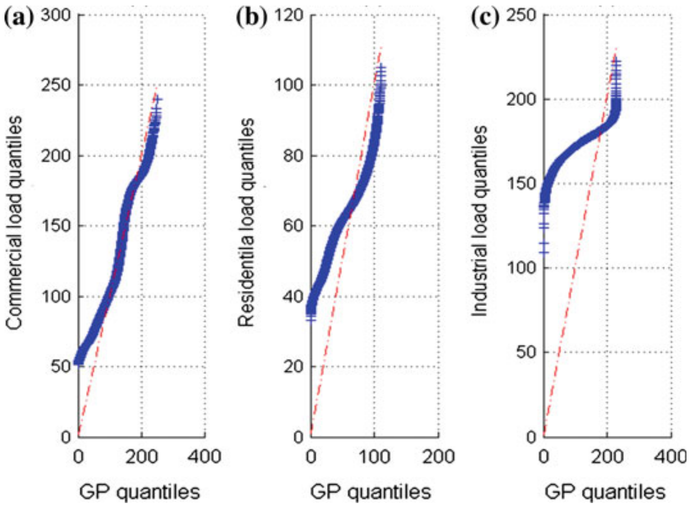


Fig. 45.8 **a** Q–Q plot of commercial load versus generalized pareto distribution, **b** Q–Q plot of residential load versus generalized pareto distribution, **c** Q–Q plot of industrial load versus generalized pareto distribution

Generalized extreme value (GEV) distribution has been used in Fig. 45.9. Unlike the previous ones, the fit is very good for the industrial and the residential cases. The commercial load cannot be fitted by this type of distribution.

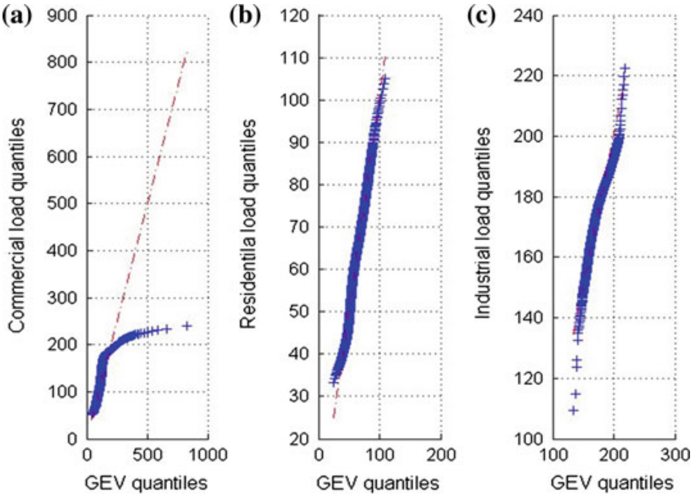


Fig. 45.9 **a** Q–Q plot of commercial load versus generalized extreme value distribution, **b** Q–Q plot of residential load versus generalized extreme value distribution, **c** Q–Q plot of industrial load versus generalized extreme value distribution

Table 45.2 Load type determination criterion

Distribution	Commercial	Residential	Industrial
R	Good	Fairly good	Bad
GP	Good	Fairly good	Bad
GEV	Bad	Good	Good

Based on the above observations a load type determination criterion can be developed. User of this criterion may plot the load versus Rayleigh, generalized Pareto and generalized extreme value distributions and compare the output with the rule of the thumb presented in Table 45.2 to find the dominant component of the load in hand. The authors of this paper would like to mention again that this criterion can be applicable only to the places where electrical heating and cooling are being used by the customers.

45.5 Results, Observations and Discussions

After pre-processing and clustering the data, thirteen sets of input variables are defined based on the available temporal and weather data as the feed for the training models. Preparing a proper set of input variables is a very significant step in any training procedure and it can strongly affect the accuracy of the method. Proper input data includes either variables with good correlation with the output data or variables that help to classify the other input variables. The input variables are year, month, day of the week, hour of the day, temperature, relative humidity, previous-day same-hour demand, previous-week same-hour demand, holidays,[6] average past-twenty-four-hour demand, average past-seven-day demand, summer temperature to help the classification of temperature in hot days, and winter temperature to distinguish the cold-day temperature.

The set of feed variables to train the models consists of thirteen column vectors of input variables with, a total of 275,513 observations in each vector and one column vector of the same size for target variables. The number of observations in each column of industrial load is 225,513.

Two different nonlinear training methods of artificial neural networks (ANN), and decision trees decision treeslearning have been used in this study. Input and target variables for the training period have been used to find the optimum configuration for ANN, and bagging decision trees. The neural network has 13 input neurons, 40 hidden neurons and one output neuron. The network trained using the back error propagation algorithm. Because of its performance under classification noise, bagging has been selected for ensembles' construction of decision trees [20]. 40 bagged regression trees have been used for the training purpose.

[6] A list of Western Australian public holidays has been used to generate the holidays input variables.

The residential and commercial models have been trained with fourteen years of data from April, 1995 to December, 2009. And, the industrial load model has been trained with eleven years of data from January 1998 to December, 2009. The testing period is the year 2010 for all the cases. Once the training procedures are done, the trained models can be used for future simulations. Using the single-day forecast and the reconstructed temperature and humidity signals as a new input set to the trained models, the forecasting horizon can be stretched from one day to seven days.

Because different benchmarks have different average load, mean absolute percentage error (MAPE) may not be able to present a good comparison. For a better comparison of the performance of the models for out of sample data (during the test year), modified mean absolute percentage error has been defined in (45.1).

$$\begin{aligned} & MMAPE_z = \frac{\frac{T}{N}\sum_{t=1}^{N} A_{zt}}{\sum_{i=1}^{T}\frac{1}{N}\sum_{t=1}^{N} A_{it}}\frac{1}{N}\sum_{t=1}^{N}\left|\frac{A_{zt}-F_{zt}}{A_{zt}}\right| \\ & A : Actual\ load \\ & F : Forecasted\ load \\ & N : Number\ of\ samples \\ & T : Number\ of\ all\ the\ regions\ except\ z \end{aligned} \tag{45.1}$$

Equation (45.1) will basically multiply the MAPE value by a coefficient which is a function of average load in different regions. The resulting MMAPE is not affected by the average load of the region itself.

Daily and weekly MMAPE of both models have been calculated for each month of the test period. The results are shown in Tables 45.3, 45.4 and 45.5. It can be seen that both methods perform very similarly. The daily MMAPE of all of the applied methods is less than 5 % which as investigated by [21] is within the range of adequate forecast and the economic impact of more accurate forecasts is very small.

Table 45.3 MMAPE of East perth for out of sample data (2010 test year)

	NN Daily MMAPE	DT Daily MMAPE	NN Weekly MMAPE	DT Weekly MMAPE	Average Temperature (°C)
Jan	2.8	3.4	3.9	4.1	24.3
Feb	2.3	2.4	3.5	3.7	24.7
Mar	2.6	2.4	3.6	3.5	22.6
Apr	2.2	1.4	3.3	2.8	18.3
May	2.3	1.4	3.3	2.9	14.2
Jun	2.1	1.5	2.9	2.7	11.2
Jul	2.3	1.8	3.1	2.9	11.8
Aug	2.3	1.6	3.3	2.7	12.2
Sep	2.1	1.4	3.0	2.8	15.1
Oct	1.9	1.3	3.0	2.5	17.5
Nov	2.1	2.5	3.3	3.8	22.0
Dec	2.3	2.3	3.3	3.6	22.6

Table 45.4 MMAPE of perth CBD for out of sample data (2010 test year)

	NN Daily MMAPE	DT Daily MMAPE	NN Weekly MMAPE	DT Weekly MMAPE	Average temperature (°C)
Jan	4.3	4.9	5.5	5.8	25.5
Feb	3.7	4.0	4.7	4.9	24.4
Mar	4.1	4.5	5.4	5.6	23.1
Apr	3.0	3.0	4.0	3.8	18.7
May	2.0	1.4	2.8	2.3	14.7
Jun	2.5	2.4	3.4	3.3	12.3
Jul	2.6	2.0	3.7	3.0	11.3
Aug	2.6	1.4	3.8	2.3	12.3
Sep	2.9	1.6	3.8	2.5	14.8
Oct	2.7	1.5	3.8	2.5	17.4
Nov	3.3	2.8	4.3	4.0	22.0
Dec	4.7	4.2	5.6	5.1	22.5

Table 45.5 MMAPE of Kalgoorlie for out of sample data (2010 test year)

	NN Daily MMAPE	DT Daily MMAPE	NN Weekly MMAPE	DT Weekly MMAPE
Jan	3.3	3.0	4.1	3.7
Feb	3.1	2.6	4.1	3.5
Mar	3.0	2.3	4	3.2
Apr	3.1	2.6	4	3.4
May	3.9	2.1	4.8	3.1
Jun	4.5	2.3	6.5	3.2
Jul	3.6	2.3	4.7	3.9
Aug	3.9	2.4	5.6	3.3
Sep	3.3	2.6	4.2	3.8
Oct	3.8	2.0	4.8	3.0
Nov	4.1	2.5	5.3	3.1
Dec	3.7	2.5	4.8	3.6

It can be observed that for hot months neural networks is a better choice for forecasting the electricity load of residential and commercial load. On the other hand it would be better to use decision trees for the rest of year. The situation is totally different for the industrial case. It can be seen that decision trees perform better for all the months of the test year of the industrial case irrespective of the temperature. Decision trees perform better than neural networks when the system's nonlinearity is low. But when the system's nonlinearity increases, neural networks will be a better choice. In load forecasting applications the higher the temperature sensitivity, the higher the nonlinearities will be in the system.

Finally it can be concluded that both methods are capable of forecasting the electricity load with a very high accuracy, but depending on the characteristics of the case study, one of them may perform better than the other. Using the introduced load type determination criterion will help the planner to extract the dominant component of the electricity load and help him/her to decide which method to use. This study suggests bagging decision trees for dominantly industrial loads. Based on the temperature sensitivity of the system decision trees or a combination of decision trees and neural networks can be used for dominantly commercial and residential cases.

References

1. Weron R (2006) Modeling and forecasting electricity loads and prices: a statistical approach. Wiley, London, p 178
2. Temraz HK, Salama MMA, Chikhani AY (1997) Review of electric load forecasting methods. In: IEEE Canadian conference on electrical and computer engineering, vol 1. pp 289–292
3. Kourtis G, Hadjipaschalis I, Poullikkas A (2011) An overview of load demand and price forecasting methodologies. Int J Energy Environ 2(1):123–150
4. Hahn H, Meyer-Nieberg S, Pickl S (2009) Electric load forecasting methods: tools for decision making. Eur J Oper Res 199(3):902–907
5. Sharaf A, Lie T, Gooi H (1993) A neural network based short term load forecasting model. In: IEEE Canadian conference on electrical and computer engineering, pp 325–328
6. Park DC, El-Sharkawi M, Marks R, Atlas L, Damborg M (1991) Electric load forecasting using an artificial neural network. IEEE Trans Power Syst 6(2):442–449
7. Amral N, King D, Ozveren CS (2008) Application of artificial neural network for short term load forecasting. In: 43rd international universities power engineering conference, pp 1–5
8. Munkhjargal S, Manusov VZ (2004) Artificial neural network based short-term load forecasting. In: The 8th Russian-Korean international symposium on science and technology, vol 1. pp 262–264
9. Peng T, Hubele N, Karady G (1993) An adaptive neural network approach to one-week ahead load forecasting. IEEE Trans Power Syst 8(3):1195–1203
10. Bala PK (2010) Decision tree based demand forecasts for improving inventory performance. In: IEEE international conference on industrial engineering and engineering management, pp 1926–1930
11. Lobato E, Ugedo A, Rouco R (2006) Decision trees applied to spanish power systems applications. In: IEEE 9th international conference on probabilistic methods applied to power systems, pp 1–6
12. Willis HL (2002) Spatial electric load forecasting, vol 71(2), 2nd edn. CRC Press, Boca Raton, p 760
13. Moghaddam M, Bahri P (2012) A novel approach for forecasting of residential, commercial and industrial electricity loads. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science, WCECS, vol. 2. pp 1365–1371
14. Asber D, Lefebvre S, Asber J, Saad M, Desbiens C (2007) Non-parametric short-term load forecasting. Int J Electr Power Energy Syst 29(8):630–635
15. Seppälä A (1996) Load research and load estimation in electricity distribution. Technical research centre of Finland
16. Hyndman RJ, Fan S (2010) Density Forecasting for Long-Term Peak Electricity Demand. IEEE Trans Power Syst 25(2):1142–1153
17. Chicco G, Napoli R, Piglione F (2001) Load pattern clustering for short-term load forecasting of anomalous days. In: IEEE porto power tech proceedings, vol 2
18. Jain A, Satish B (2009) Clustering based short term load forecasting using artificial neural network. In: IEEE/PES power systems conference and exposition, pp 1–7

19. Meng M, Chang Lu J, Sun W (2006) Short-term load forecasting based on ant colony clustering and improved BP neural networks. In: International conference on machine learning and cybernetics, pp 3012–3015
20. Dietterich T (2000) An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach Learn 22:1–22
21. Ranaweera DK, Karady GG, Farmer RG (1997) Economic impact analysis of load forecasting. IEEE Trans Power Syst 12(3):1388–1392

Chapter 46
Energy Consumption of Biodiesel Production from Microalgae Oil Using Homogeneous and Heterogeneous Catalyst

Nezihe Azcan and Ozlem Yilmaz

Abstract Transesterification reaction of the new generating biofuel resource, microalgae was conducted. In order to evaluate the effects of reaction variables such as catalyst type (KOH, NaOH), catalyst amount (1–1.5 % w/w), oil:methanol molar ratio (1:6–1:10), and time (5–20 min) on the methyl ester content of biodiesel. Suitable transesterification reaction conditions were determined as 65 °C, 1 wt % catalyst amount, 5 min, 1:8 microalgae oil:methanol molar ratio using microwave heating system for both homogeneous catalysts. At these conditions fatty acid methyl ester content was determined as 96.54 and 96.82 % for KOH and NaOH, respectively. The results show that microwave heating has effectively reduced the reaction time from 210 min (for conventional heating) to 5 min. Electric energy consumption for microwave heating in this accelerated transesterification reaction was only 28.22 % of estimated minimum heat energy demand. Heterogeneous base catalyst, KOH/Al_2O_3, was also used to investigate the reaction activity due to growing interest in transesterification. According to slow reaction rate of heterogeneously catalyzed reactions, time, oil:methanol molar ratio and catalyst amount were increased as 35 min, 1:12 and 3 % w/w, respectively. Biodiesel conversion was found to be 97.79 % at these conditions.

Keywords Base catalyst · Biodiesel · Energy consumption · Heterogeneous catalyst · Homogeneous catalyst · Microalgae oil · Microwave irradiation · Transesterification

N. Azcan (✉) · O. Yilmaz
Faculty of Engineering, Department of Chemical Engineering, Anadolu University,
2 Eylul Campus 26470 Eskişehir, Turkey
e-mail: nazcan@anadolu.edu.tr

O. Yilmaz
e-mail: oyilmaz@anadolu.edu.tr

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_46,

Abbreviation

TR	Temperature rise
TM	Temperature maintenance
h	Heat transfer coefficient, Wm^{2}K^{-1}
A_s	Surface area, m^2
m	Mass of the materials, kg
c	Specific heat of the materials, Jkg^{-1}K^{-1}
Ti	Initial temperature, K
T_{amb}	Ambient temperature of the surrounding air, K
T_{emp}	Temperature
t	Time
hh	Hour
mm	Minute
ss	Second
tan δ	Tangent loss factor

46.1 Introduction

Since fossil fuel depletion, fluctuation in crude oil prices, increasing environment and energy interests have demand for renewable fuels such as biodiesel which is obtained from the oils and fats. Biodiesel is a renewable, biodegradable, non-toxic biofuel, and essentially free of sulfur and aromatics that shows great promise. It is derived from the transesterification of mono-, di- and tri-acylglycerides (TAGs) [1, 2]. Biodiesel can be obtained by transesterification reaction of any fats and oils feedstock like oil-bearing crops, animal fats, and algal lipids (Fig. 46.1). The literature contains 100 of references of biodiesel production from a wide variety of feedstock's [2, 3]. Raw material availability of raw vegetable oil has been recently a critical issue for the biodiesel production. Most of biodiesel is produced from edible vegetable oils. There is concern that at anticipated future production levels of the use of edible oils for fuel production will compete significantly with food uses. This would result in an undesirable increase in food and biodiesel costs, a particularly damaging occurrence in the case of biodiesel, which, even at current prices, cannot compete with petroleum fuel on an economic basis.

Algae are theoretically very promising source of biodiesel. Currently considerable attention has been focused on production of biofuels from algal biomass [4]. Microalgae can provide renewable biofuels in several different ways including photobiological biohydrogen production, methane production derived from anaerobic digestion of the algal biomass, and biodiesel produced from microalgal oil [5]. The microalgae oil for biodiesel production has some advantages which were listed below.

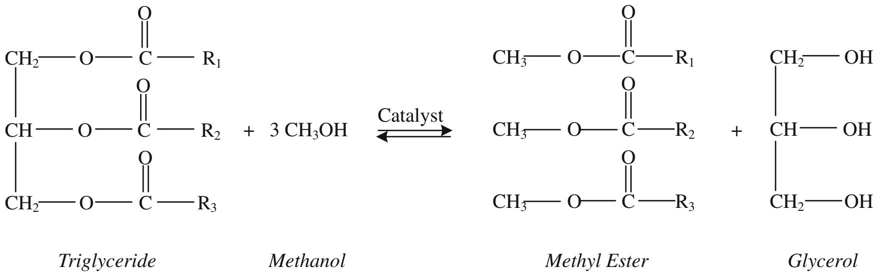


Fig. 46.1 Transesterification reaction of biodiesel

- Benefits of algae include high yields and flexibility of location and feedstock [6],
- Produce higher yields of oil per hectare of land,
- Do not require arable land or fresh water to grow and thus do not complete with food crops [7],
- High levels of polyunsaturates in algae biodiesel is suitable for cold weather climates [8].

Microalgae can produce both biofuels and valuable co-products such as omega-3 and several metabolites of economic interest, such as carotenoids (e.g., astaxanthin, lutein), vitamin E (alpha–tocopherol), and polyunsaturated fatty acids (arachidonic and γ linoleic). Thus, they have a large potential for feed, food, cosmetics, and pharmaceutical industries, which makes their conversion to biodiesel cost efficient [9, 10].

Various methods have already been used for biodiesel production from microalgae biomass. A process containing of extraction followed by transesterification was performed. Lipid extraction consists of different techniques such as Soxhlet extraction [11], ultrasonification (chloroform, methanol mixture, methanol) [12], supercritical fluid extraction [9]. Transesterification reaction was performed in the presence of different catalyst such as acid [13], alkali and enzymatic [9, 12].

Microwave activation is now widely accepted and a very popular non-conventional alternative technology in organic chemistry, and often displays improvement over the conventional heating [14]. When the reaction is carried out under microwaves, transesterification is efficiently accelerated in a short reaction time [15]. The use of microwave heating excludes many heat-transfer processes and thus energy losses that are present with conventional oil-bath heating. Moreover, the direct heating under microwave irradiation provides a more homogeneous heat profile when compared to oil-bath heating where the vessel walls are significantly warmer than the contents [16].

Homogeneous catalysts like KOH, NaOH have the advantages of high activity and mild reaction conditions. However, the separation of the catalyst in the reaction medium after transesterification is difficult and a large amount of waste water is used in order to obtain clean biodiesel.

Development of heterogeneous catalysts is an alternative technique that could eliminate the additional running costs of separation and purification [17].

Microalgae oil was used in this study in order to obtain biodiesel by the transesterification reaction under microwave irradiation which is effective in accelerating the reaction rate, conventional heating and at room remperature using homogenous and heteregenous base catalysts. Energy consumption of each method was individually calculated and compared. Some of datas were published in the proceedings as lecture notes [18].

46.2 Experimentals and Methods

46.2.1 Materials

Microalgae oil, *Chlorella protothecoides*, was provided by Soley Institute (Istanbul, Turkey). Methanol, potassium hydroxide, sodium hydroxide, pyridine and N-methyl-N-trimethysilyltrifluoroacetamide (MSTFA) were supplied from Sigma-Aldrich. Standards of fatty acid methyl esters were purchased from Accu Standards. KOH/Al_2O_3 catalyst with the loading of 30 % w/w, synthesized by wet-impregnation method, was used [19].

46.2.2 Equipment

Microwave assisted transesterification experiments were conducted at Start S model microwave unit, supplied from Milestone Company-Italy. The system (Fig. 46.2) was equipped with a reflux condenser, a magnetic stirrer bar and a

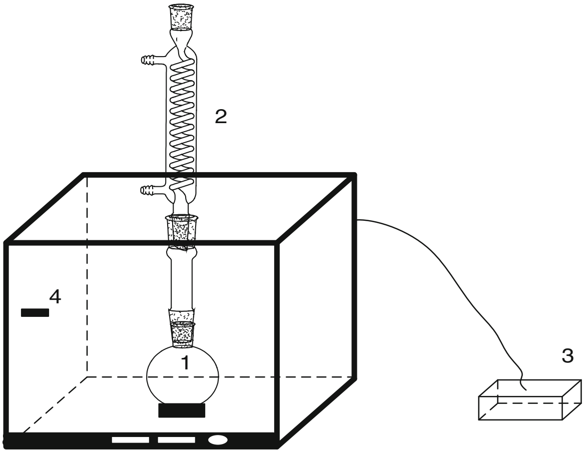


Fig. 46.2 Microwave synthesis unit; (*1*) glass reactor; (*2*) condenser; (*3*) control unit; (*4*) infrared temperature sensor

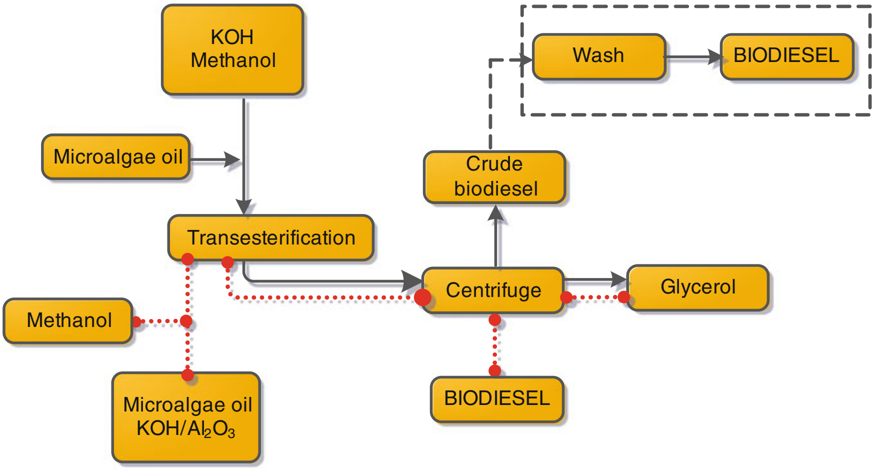


Fig. 46.3 Experimental procedure in order to obtain biodiesel

non-contact infrared continuous feedback temperature system which allows continuous stirring and constant temperature control.

Fatty acid composition of microalgae oil and methyl ester content of biodiesel were determined by Agilent N Gas Chromatography apparatus equipped with flame ionization detector.

46.2.3 *Experimental Procedure*

Transesterification reactions were carried out using heating system (microwave synthesis unit, water bath as conventional heating) and at room temperature. Scheme of the transesterification reaction carried out using homogeneous (KOH, NaOH) and heterogeneous (KOH/Al_2O_3) catalysts are given in Fig. 46.3.

46.2.3.1 Transesterification Reaction Assisted by Microwave Heating

Transesterification reactions were conducted at microwave synthesis unit in the presence of homogeneous and heterogeneous catalysts. As a homogeneous catalyst KOH and NaOH were used at 65 °C temperature (Table 46.1) and KOH/Al_2O_3 was used as heterogeneous catalyst at various reaction time (5–35 min), 65 °C, using 1:12 oil:methanol molar ratio and 3wt. % catalyst amount.

Table 46.1 Reaction conditions (KOH and NaOH)

Time (min)	Oil:methanol molar ratio	Catalyst amount (wt %)
5	1:6	1.0
10	1:8	1.5
15	1:10	–
20	–	–

46.2.3.2 Reaction at Room Temperature

Transesterification reactions were performed at 25 °C, using various oil:methanol molar ratio (1:6, 1:8, 1:10) and time (60, 120, 180, 240, 300 min) in the presence of 1wt. % KOH and NaOH.

Magnetic stirrer bar was used during the reaction.

46.2.3.3 Conventional Heating System

Transesterification reactions were done at 65 °C, 1:8 oil:methanol molar ratio 210 min in the presence of 1wt. % KOH and NaOH.

46.2.4 Analytical Methods

46.2.4.1 Determination of Physicochemical Properties of Microalgae Oil

Relative fatty acid composition of microalgae oil was determined as methyl esters of fatty acid by gas chromatography analysis using Agilent 6,890 N gas chromatography apparatus with HP-Innowax column (60 mL × 0.25 mm ID × 0.25 μm film thickness) after converting fatty acids into methyl ester forms using 14 % BF_3 in methanol [20]. A necessary procedure associated with this analysis is lipid derivatization. This process changes the volatility of lipid components, and improves peak shape and thus provides better separation [21]. Helium was used as a carrier gas at a flow rate of 1.0 mL/min. Temperature program was started at 60 °C, heated at 4 °C/min to 220 °C and heated to 240 °C at 1 °C/min, staying at this temperature for 10 min [15, 22].

Relative density, viscosity, saponification number, acid value, peroxide value and iodine value of microalgae oil were determined according to standard methods [20, 23].

46.2.4.2 Determination of Ester Content

Biodiesel purity is defined as the methyl ester content of biodiesel. The methyl esters were firstly derivatized by N-methyl-N-trimethysilyltrifluoroacetamide (MSTFA) at 25 °C, for 15 min which is known as silylation. It is the most common method used to derivate organic compounds containing active hydrogen atoms (e.g. –OH, = NH, $-NH_2$, –SH and –COOH), which results in products with reduced polarity, enhanced volatility and increased thermal and catalytic stability necessary for optimal sensitivity and resolution of various components in mixtures by GC analyses. Therefore the methyl esters obtained from the transesterification of microalgae oil were reacted with MSTFA at 25 °C, for 15 min. After silylation, the methyl esters, MG, DG, TG and ester content were identified using gas chromatography (Agilent 6,890 N) equipped with DB-5HT column (15 m × 0.32 mm ID × 0.10 μm film thickness) and flame ionization detector. Temperature program was started at 50 °C, heated at 15 °C/min to 180 °C and heated to 230 °C at 7 °C/min, then heated at 10 °C/min to 370 °C, staying at this temperature for 20 min [15, 22].

Fuel Properties of Biodiesel

Physical properties of biodiesel such as relative density, viscosity, flash point, heating value and ester content were determined using standard test methods according to EN 14214.

Estimation of Minimum Heat Energy Consumption

The minimum heat energy consumption during maintaining reaction temperature could be defined to heat dissipation of the reactor to the surrounding air. The heat dissipation was estimated by the lumped capacitance method which may be used to determine the variation of the temperature with time [24–26]. If there is no external heat supply, because no temperature gradient exists on the material mixture due to strong stirring, transient temperature variation is determined by the heat dissipation as following,

$$-hA_s(T - T_{amb}) = mc\, dT/dt \tag{46.1}$$

The left side of Eq. 46.1 is the heat dissipation of the reactor. If the constant hA_s is obtained, the heat dissipation can be obtained as the temperatures of reaction and surrounding air. From Eq. 46.1, transient temperature variation becomes.

$$T = (T_i - T_{amb})\exp[-hA_s/mct] + T_{amb} \tag{46.2}$$

The constant hA_s could be estimated by fitting the experimental result to the Eq. 46.2 [24].

46.3 Results and Discussion

The fatty acid composition of the microalgae oil was determined by gas chromatography. It was found that the fatty acids of the oil were composed primarily of 65.39 % oleic acid, 20.89 % linoleic acid, 6.23 % linolenic acid, 4.99 % palmitic acid, 1.66 % stearic acid, 0.55 % arachidic acid, 0.23 % palmitoleic acid, 0.15 % lauric acid. Fatty acid composition of microalga is in good agreement with the literature [27]. The molecular weight of the oil was calculated as 880.32 g/mol according to fatty acid compositions.

The physicochemical properties of microalgae oil (Table 46.2) are in the range of the literature [27].

Biodiesel conversion (methyl ester content) was determined under different reaction time, temperature, catalyst type and loading, oil:methanol molar ratio and heating system in order to minimize the reaction time with a maximum conversion.

46.3.1 Microwave Heating

Two catalyst:oil ratios (1.0 and 1.5 wt %) were used at 65 °C, 1:6 oil:methanol molar ratio, 15 min. 1 wt % KOH and 1 wt % NaOH gave the highest biodiesel purity as 96.25, 96.40 %, respectively. So that in further experiments 1.0 wt % catalysts was used. Combined effect of time and oil:methanol molar ratio was investigated (Fig 46.4).

Table 46.2 Physicochemical properties of microalgae oil

Moisture (%)	Relative density	Iodine index	Acid value	Peroxide value
0.04	0.9127	115.39	0.22	11.95

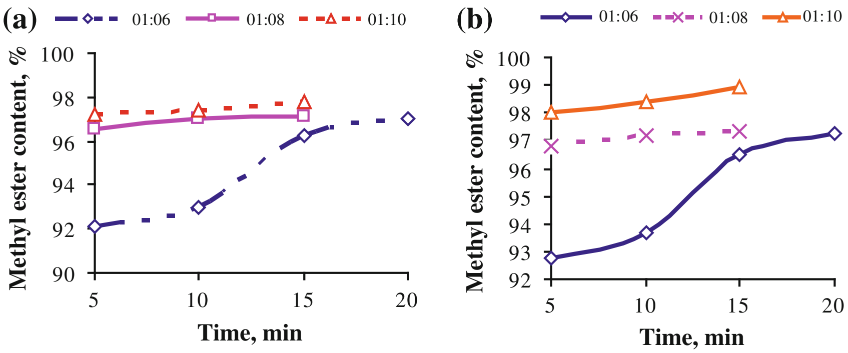


Fig. 46.4 Effect of time and oil:methanol molar ratio on biodiesel conversion

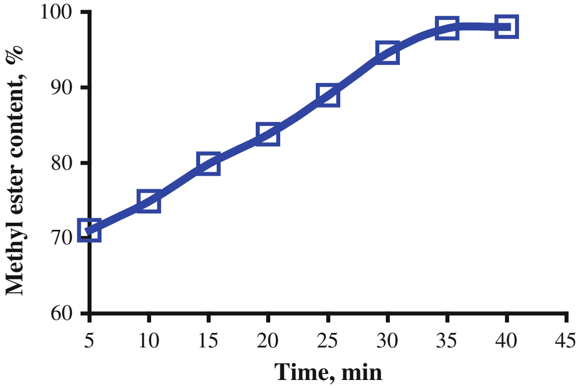


Fig. 46.5 Effect of time on biodiesel conversion in the presence of KOH/Al_2O_3

According to Fig. 46.4-a, b oil:methanol molar ratio has a significant effect on reaction time and fatty acid methyl ester content of biodiesel (conversion) for both catalyst. Over 96.50 % methyl ester content (which is in the range of EN 14214) of biodiesel was obtained using 1:6 oil:methanol molar ratio after 20 min reaction time. Similar purity was achieved after 5 min using 1:8 oil:methanol molar ratio for both catalysts.

Results of heterogeneous catalyst are shown in Fig. 46.4.

97.79 % methyl ester content of biodiesel was achieved at 35 min using the heterogeneous catalyst (Fig. 46.5).

46.3.2 *Room Temperature Reaction*

Transesterification reactions were done at room temperature to compare with the results obtained by microwave heating. Reactions were performed at different time (60–300 min) and oil:methanol molar ratio (1:6–1:10) at 25 °C, 1 %wt. KOH and NaOH. Experimental results are shown in Fig. 46.6.

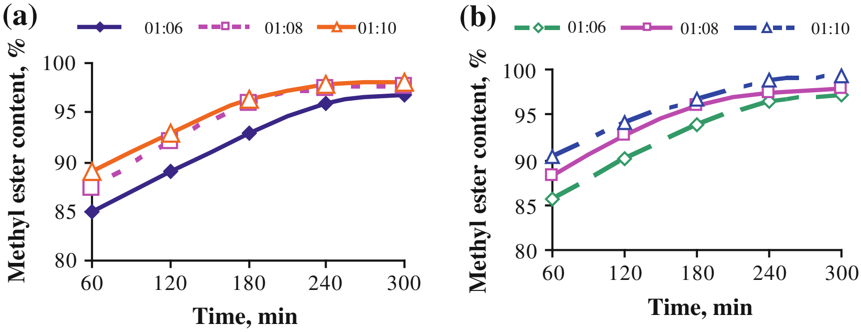


Fig. 46.6 Effect of time and oil:methanol molar ratio on biodiesel conversion

Table 46.3 Fuel properties of resulting biodiesel from microalgae oil

Properties	Biodiesel	Standard values
Relative density	0.8772	0.86–0.90
Viscosity (mm^2/s, 40 °C)	4.51	1.9–6.0
Moisture content (%)	0.08	0.05 (max)
Pour point (^{o}C)	−20	0 (max)
Cloud point (^{o}C)	−4	–
Heating value, J/g	38,339	35,000–40,000

As it is seen in Fig. 46.6a–b, biodiesel conversion increases with time up to 240 min then keep constant for both homogeneous catalysts. Conversion increases with the oil:methanol molar ratio up to 1:8 which is almost the same with 1:10. Conversions were obtained as 97.34 % (KOH) and 97.47 % (NaOH) at 1:8 oil:methanol molar ratio and 240 min reaction time.

46.3.3 Conventional Heating

96.94 and 97.32 % methyl ester content was obtained using 1:8 oil:methanol molar ratio at 65 °C and 210 min reaction time at 1wt. % KOH and 1wt. % NaOH, respectively.

46.3.4 Fuel Properties of Biodiesel

Comparison of fuel properties of microalgae oil with standard value are given in Table 46.3. Obtained values are coherent with EN 14214.

46.3.5 Microwave Heating Consumption

Microwave energy consumption was calculated according to method used by Kim et al. [24]. The absorbed microwave power is shown in Fig. 46.7.

The energy consumptions were obtained in separated two stages of temperature rise and maintenance for convenience as shown in Table 46.4.

Total energy consumption could be determined from the reaction time of the microwave heating. The microwave power for maintaining temperature was averaged of the absorbed microwave power during the reaction [24]. The energy of microwaves comes from electrical energy that is converted by a power supply to high voltages that in turn are applied to the microwave power tube or generator [28]. Therefore, the energy consumption of electricity should be compared with the one of the conventional heating in order to show the energy-efficiency.

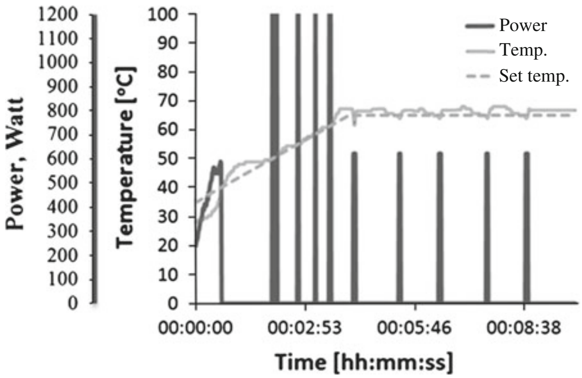


Fig. 46.7 Absorbed microwave power and temperature profile (65 °C, 1 wt % KOH, 1:8 oil:methanol molar ratio, 5 min)

The electric energy consumption was calculated by dividing the measured microwave energy consumption by 0.60 considering typical energy conversion rates of high-voltage power supply and the magnetron to 0.80 and 0.75, respectively [24].

The estimated minimum heat energy demand for temperature rise was calculated by heat capacities of the reactor composed of oil of 29.9922 g, methanol of 9 g, catalyst of 0.2961 g, and glass vessel of 90.2528 g. The other energy consumptions such as evaporation of methanol, reaction energy and heat loss did not consider here. The minimum heat demand for the temperature maintenance was obtained from a measurement of heat dissipation from the reactor to the surrounding air. Heat transfer constant of hA_s was determined as 0.1708 W/K using Eq. 46.2. The heat dissipation rate (minimum heat energy consumption) at the reaction temperature was calculated to 6.77 W.

As a result, the value of microwave energy consumption during temperature rise and maintenance is higher than the estimated minimum heat energy demand because of dielectric property of reaction mixture. A material's heating rate is governed by the amount of microwave power input to it. Power level requirements are based on the properties of the material being heated for a particular throughput

Table 46.4 Energy consumption of microwave synthesis unit

	Microwave	Electricity[a]	Minimum heat demand
TR (KJ)	12.19	20.33	7.52[b]
TM (KJ)	3.53	5.87	85.30[c]
Total (KJ)[d]	15.72	26.20	92.82

[a] Electric energy (or power) consumption for microwave generation: Only 60 % electric energy converts to microwave.

[b] Estimated minimum heat energy demand calculated by heat capacities of 29.9922 g—WCO, 9 g—methanol, 0.2961 g—catalyst, 90.2528 g—glass vessel. The initial and final temperatures were 25 and 65 °C, respectively.

[c] Heat loss of the reactor to the surrounding air calculated by the measured heat transfer coefficient of the reactor at 65 °C with the ambient temperature of 25 °C.

[d] Energy consumption until conversion of more than 96.5 % could be obtained; the reaction times of 5 min for microwave heating and 210 min for the conventional heating were used.

and the initial and final temperatures [29]. Another reason can be caused by polarization which takes place when the effective current in the irradiated sample is out of phase with that of the applied field by a difference (termed δ). This difference defines the tangent loss factor, tan δ, often named the dissipation factor or the dielectric loss tangent. The word “loss” refers to the input microwave energy that is lost to the sample by being dissipated as heat [30].

Compared to the conventional heating method, microwave irradiation can reduce reaction time and save energy significantly due to the fast and volumetric heating effect [31]. Reaction time changes according to heating system in order to reach 96.5 % conversion such as 5, 210, 240 min for microwave, conventional and room temperature, respectively.

Eventually, the electric energy consumption becomes only 28.22 % of the estimated minimum heat demand Table 46.4.

Magnetic stirrer consumes electrical energy to mix the reactants and 630 kJ was used for 240 min.

46.4 Conclusion and Future Work

Transesterification reaction of microalgae oil were carried out in the presence of homogeneous (KOH, NaOH) and heterogeneous (KOH/Al_2O_3) catalysts. Effect of heating system on the biodiesel conversion was determined using homogeneous catalyst. NaOH gave the highest biodiesel conversion 96.82 % 1:8 oil:methanol at 65 °C and 5 min reaction time by using microwave heating system. 97.79 % conversion was obtained using KOH/Al_2O_3 at 65 °C, 1:12 oil:methanol molar ratio, 3wt. % catalyst amount and 35 min reaction time. Although reaction time is significantly higher than homogeneous catalyst, heterogeneous catalysts have the general advantage of being reusable and easy to separate from the reaction products. Fuel properties of biodiesel comprise with the standard values.

While there are many heating techniques such as conventional, ultrasound, water-bath; they require longer reaction times, and are energy- and cost-intensive. Microwave heating shows a promising technique for biodiesel production in a short reaction time with high product yield. Electric energy consumption for microwave heating in this accelerated transesterification reaction was only 28.22 % of estimated minimum heat energy demand because of significantly reduced reaction time.

References

1. Kim HJ, Kang BS, Kim MJ, Park YM, Kim DK, Lee JS, Lee KY (2004) Transesterification of vegetable oil to biodiesel using heterogeneous base catalyst. Catalysis Today 93–95:315–320
2. Krohn BJ, McNeff CV, Yan BW, Nowlan D (2011) Production of algae-based biodiesel using the continuous catalytic Mcgyan® process. Bioresour Technol 102:94–100

3. Hoekman SK, Broch A, Robbins C, Ceniceros E, Natarajan M A (2012) Review of biodiesel composition, properties, and specifications. Renew Sustain Energy Rev 16(1):143–169
4. Demirbas A (2011) Biodiesel from oilgae, biofixation of carbon dioxide by microalgae: A solution to pollution problems. Appl Energy 88(10):3541–3547
5. Pai TY, Lai WJ (2011) Analyzing algae growth and oil production in a batch reactor under high nitrogen and phosphorus conditions. Int J Appl Sci Eng 9(3):161–168
6. Phyper JD, MacLean P (2009) Good to green: managing business risks and opportunities in the age of environmental awareness. Wiley, Canada
7. Schmidt M (2012) Synthetic biology industrial and environmental applications. Wiley, Germany, p 36
8. Gonzalez-Delgado AD, Kafarov V (2011) Microalgae based biorefinery: issues to consider. CT&F - Ciencia, Tecnología 4(4):5–21
9. Koberg M, Cohen M, Ben-Amotz A, Gedanken A (2011) Bio-diesel production directly from the microalgae biomass of Nannochloropsis by microwave and ultrasound radiation. Bioresour Technol 102:4265–4269
10. Sivakumar G, Xu J, Thompson RW, Yang Y, Randol-Smithd P, Weathers PJ (2012) Integrated green algal technology for bioremediation and biofuel. Bioresour Technol 107:1–9
11. D’Oca MGM, Viêgas CV, Lemões JS, Miyasaki EK, Morón-Villarreyes JA, Primel EG (2011) Abreu PC Production of FAMEs from several microalgal lipidic extracts and direct transesterification of the Chlorella pyrenoidosa. Biomass Bioenergy 35:1533–1538
12. Tran DT, Yeh KL, Chen CL Chang JS (2012) Enzymatic transesterification of microalgal oil from Chlorella vulgaris ESP-31 for biodiesel synthesis using immobilized Burkholderia lipase. Bioresour Technol 108:119–127
13. Ehimen EA, Sun ZF, Carrington CG (2010) Variables affecting the in situ transesterification of microalgae lipids. Fuel 89:677–684
14. Loupy A, Varma RS (2006) Microwave effects in organic synthesis Mechanistic and reaction medium considerations. Chemistry Today 24(3):36–40
15. Azcan N, Danisman A (2008) Microwave assisted transesterification of rapeseed oil. Fuel 87:1781–1788
16. Hoogenboom R, Wiesbrock F, Schubert U (2006) Microwave—assisted polymerization: the living cationic ring-opening polymerization of 2-oxazolines. Chemistry Today 24(3):46–49
17. Romero R, Martinez SL, Natividad R (2011) Biodiesel production by using heterogeneous catalysis. In: Manzanera M (ed) Alternative Fuel. ISBN: 978-953-307-372-9
18. Azcan N, Yilmaz O (2012) Microwave irradiation application in biodiesel production from promising biodiesel feedstock: microalgae (*Chlorella protothecoides*). Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, WCECS 2012, 24–26 Octo, 2012, San Francisco, pp 737–742
19. Azcan P, Misiroglu (2012) Activity of heterogenous catalyst for biodiesel production, proceedings of 20th European biomass conference and exhibition 2012, EU BC&E 2012, Milan, pp 1776–1777
20. Helrich K Official methods of analysis of the association of official analytical chemists, (15th edn). Association of official Analytical Chemists, Inc, Arlington
21. Liu KS (1994) Preparation of fatty acid methyl ester for gas-chromatographic analysis of lipids in biological materials. JAOCS 71:1179–1187
22. Azcan N, Danisman A (2007) Alkali catalyzed transesterification of cottonseed oil by microwave irradiation. Fuel 86:2639–2644
23. The United States Pharmacopoeia (U.S.P. XXII) (1990) Mark Printing Co, Easton, USA
24. Kim D, Choi J, Kim GJ, Seol SK, Ha YC, Vijayan M, Jung S, Kim BH, Lee GD, Park SS (2011) Microwave-accelerated energy-efficient esterification of free fatty acid with a heterogenous catalyst, Bioresour Technol, 102:3639–3641
25. Incropera FP, Dewit DP, Bergman TL, Lavine AS (2007) Fundamental heat and mass transfer, (6th edn). John Wiley, New York, pp 256–258
26. Bergman TL, Lavine AS, Incropera FP, Dewitt DP (2011) Introduction to heat transfer, (6th edn). John Wiley & Sons, New York, p 280

27. Chen YH, Huang BY, Chiang TH, Tang TC (2012) Fuel properties of microalgae (Chlorella protothecoides) oil biodiesel and its blends with petroleum diesel. Fuel 94:270–273
28. Hui YH (1992) Encyclopedia of food science and technology, (3rd edn). Wiley, p 770
29. Techcommentary industrial microwave heating applications (1993). EPRI Center for Materials Fabrication 4(3), http://infohouse.p2ric.org/ref/10/09077.pdf
30. Fernández Y, Arenillas A, Menéndez JÁ (2011) Microwave heating applied to pyrolysis. In: Grundas S (ed) Advances in induction and microwave heating of mineral and organic materials InTech, pp 728–729
31. Ma L, Chen WX, Zhao J, Zheng YF (2007) Synthesis of $Pr(OH)_3$ and Pr_6O_{11} nanorods by microwave-assisted method: effects of concentration of alkali and microwave heating time. J Cryst Growth 303:590–596

Chapter 47
Detrended Fluctuation Analysis: An Experiment About the Neural-Regulation of the Heart and Motor Vibration

Toru Yazawa and Yukio Shimoda

Abstract Acceleratory and inhibitory cardio-regulator nerves innervate the heart of a living creature. The two nerves discharge concurrently to maintain an equilibrium state of the heart. The nerves change their frequency of discharge in a reflexive manner to meet the demand from the periphery, such as augmentation of oxygen supply or vice versa. Consequently, the heart exhibits a dynamic change in rate of pumping and force of contraction. If the control system fails, the heart exhibits an unhealthy state. However, an assessment of a healthy/unhealthy status is uneasy, because we are not able to monitor the nerve activities by non-invasive methods. Therefore, we challenged to detect a state of the heart without nerve-recordings. We used the Detrended Fluctuation Analysis (DFA) by applying it to a heartbeat interval time series because the DFA is believed that it can quantify the state of heart. The objective of this research was to determine whether the DFA technology could function as a useful method for the evaluation of the subject's quality of a cardiovascular-related illness. We performed DFA on the EKGs (Electrocardiograms) from living organisms and a running motor as well. We conclude that scaling exponents could determine whether the subjects are under sick or healthy conditions.

Keywords Animal model · Cardio-vascular system · Crustacean heart · DFA · Fluctuation analysis · Heartbeat · Interval · Motor vibration · Scaling exponent · Time series

T. Yazawa (✉)
Neurobiology, Bio-Physical Cardiology Research Group, Biological Science, Tokyo Metropolitan University, 1-1 Minami Ohsawa, Hachioji, Tokyo 192-0397, Japan
e-mail: yazawa-tohru@tmu.ac.jp

Y. Shimoda
Medical Research Institute, Tokyo Women's Medical University, Shinjuku, Tokyo 162-0054, Japan
e-mail: yshimoda@lab.twmu.ac.jp

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_47,

47.1 Introduction

Despite the development in the field of heart disease with pharmacotherapy and a device for resynchronization therapy, the number of hospitalizations for heart failure in the United States each year exceeds over 1 million, and the mortality still remains high [1]. Technology is required for much more improvement of our ability to issue early warnings. However, there is no straightforward theory that can predict when a heart failure might occur. We cannot hope to improve public health without a shift into early detection and prevention of a disease. The key question is how to make an early detection. We propose that a computation method on a heartbeat-interval time series is practically useful for distinguishing between a sick heart and a healthy heart. The Detrended Fluctuation Analysis (DFA) was originally developed by Peng et al. [2, 3], to check power-law characteristics of the heartbeats. Since Peng's publication, it has been widely accepted that a healthy heart exhibits a healthy scaling-exponent, which is one (1.0). We here show results of DFA obtained from various living organisms, including humans. The present tests revealed that DFA could describe brain–heart interactions quantitatively. We conclude that scaling exponents could determine whether the subjects are under sick or healthy conditions on the basis of cardiac physiology. We believe that DFA is a new, useful numerical method for quantifying the degree of wellness and the transition from illness to wellness and vice versa.

47.2 Materials and Methods

47.2.1 Peak Detection

Our heartbeat-interval analysis requires detection of the precise timing of the heartbeat. A consecutive and perfect detection without missing any beat is necessary. According to our preliminary tests, about 2,000 consecutive heartbeats were required for obtaining a reliable computation of scaling exponent. Peng et al. [2] suggested that, in his e-mail to the author, longer recording of the heartbeats would give better results. However, we found that a long recording was not justifiably useful and a recording of about 2,000 consecutive heartbeats are preferable for a practical use.

To detect the timing of the heartbeats, one may assume that a common Electrocardiograms (EKGs) recording is sufficiently useful. However, the problem with a conventional EKG was the drifting of the baseline of the recording. Due to the drift and the contamination of an unexpected noise, recording failures may happen.

Another obstacle arose from the premature ventricular contraction (PVC). Among the "normal" subjects (age over 40 years old), about 60 % of subjects have PVC arrhythmic heartbeats. Normally, this PVC is believed to be a benign arrhythmia, and in fact during our recording, we found many healthy-looking individuals exhibited this arrhythmia. However, PVC is an obstacle to a perfect detection of accurate timing of the heartbeat, because the height of its signal varies often. If the baseline of EKG recording could be stable, the heartbeats would automatically be detectable, even when irregular beats appeared sporadically. Unfortunately, in commercial EKG recording devices, the baseline of the recording is not stable.

47.2.2 Stable Baseline

To capture heart beat peaks without missing any detection, we made an EKG amplifier that stabilizes the baseline of the recording (Fig. 47.1). The important issue was: we discovered that a time-constant for an input-stage of recording must be adjusted to an appropriate level (the ideal time would be, $\tau < 0.22$ s).

Having a stable baseline recording was an advantage to our DFA research. However, in some cases, inevitable noises would ruin the recording. In such case, we removed the noises by identifying them visually on the PC screen, thus making a perfect (without miscounting) heartbeat-interval time series. We have already identified how this inconvenience occurred. Most of these cases were due to the sweat on the skin under the electrodes. We were able to overcome this problem by cleaning the skin with an appropriate solution.

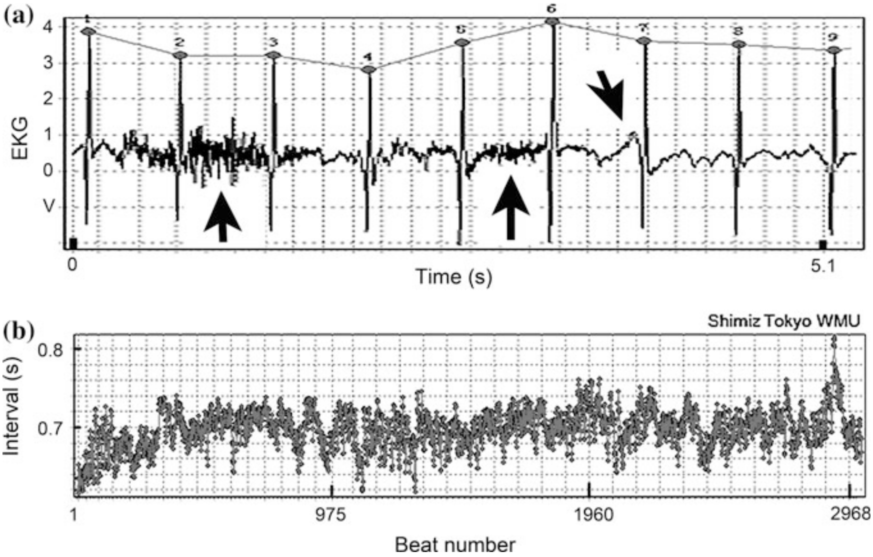


Fig. 47.1 **a** An example of the baseline-stable recording. One can see nine peaks that were captured automatically. Subject moved (pointed by *arrows*). **b** A heartbeat-interval time series from the recording-**a**, about 3,000 beats, was made at once

47.2.3 DFA: Background

DFA is based on the concepts of "scaling" and "self-similarity" [4]. It can identify "critical" phenomena, because the systems near critical points exhibit self-similar fluctuations [2, 5], which means that recorded signals and their magnified/contracted copies are statistically similar. In general, statistical quantities, such as "average" and "variance", of fluctuating signals can be calculated by taking the average of the signals through a certain section; however, the average is not necessarily a simple average. In this study, we took a squared average of the data. The statistical quantity calculation depended on the section size. The scaling exponent and DFA are well explained recently by Stadnitsuki [4]: Consult the article about fractal, scaling, Hurst-exponent, and power spectral density, regarding to fractality research [4]. Here, we used α as the "scaling exponent", which characterizes self-similarity.

Stanley and colleagues considered that a scaling property can be detected in biological systems, because most of these systems are strongly nonlinear and resemble the systems in nature, which exhibit critical phenomena. They applied the DFA to DNA arrangement and EKG data and discovered the usefulness of the scaling property [2, 6], and emphasized the potential utility of DFA in life sciences [6]. Although the practical medical use of DFA technology has not progressed to a great extent, nonlinear technology is now widely accepted [4], and rapid advances are being made in this technology.

47.2.4 DFA: Technique

We made our own computation program, based on the previous publication [2–4], which is described in one of the references [7].

47.2.5 Heartbeat Recording

For heartbeat recordings, we used a Power Lab System (AD Instruments, Australia). For EKG electrodes, a set of ready-made three AgAgCl electrodes (+, −, and ground; Nihonkoden Co. Ltd. disposable Model Vitrode V) were used. Wires from EKG electrodes were connected to our newly made amplifier. These EKG signals were then connected to a Power Lab System. Finger pulse recordings were also used with a Power Lab System.

47.2.6 Volunteers and Ethics

Heartbeats were recorded outside of the hospital; university laboratory, convention hall (Innovation Japan Exhibition) etc. All subjects were treated as per the ethical control regulations of our universities, Tokyo Metropolitan University, Tokyo Women's Medical University.

47.3 Results

47.3.1 Fractal and Scaling in Biology

Numerous studies identified fractal noise in biology, including human behavior and heart physiology. According to the theory of Self Organized Criticality, a long 1/f scaling is a signature of complex dynamical systems [4]. The 1/f scaling of heartbeat time series is a typical signature of health, as shown by Kobayashi and Musha [8]. Technologically, we have confidence in this technique, with subjects, who exhibit 1/f scaling, are healthy. If the scaling exponents are not 1, the subjects are identified as unhealthy. This rule might be useful and valuable to test how it works in our biological data. To investigate 1 or not 1, we selected the method of DFA. Before that, among some popular estimators of fractal parameters, such as a spectral density and a scaling exponent, we first tried the best known method, the power spectral density (PSD), because Kobayashi and Musha used it [8].

We tested the PSD on two kinds of the lobster heart data: one was a heartbeat recorded from an isolated heart and the other one was from an intact heart. It is important to acknowledge that isolated hearts do not receive cardio-regulator nerve impulses, instead, intact hearts receive dynamical control from the cardiac center of the brain. We expected that the PSD discriminates an isolated-heart from an intact-heart. However, we found that PSD did not work well and DFA did discriminate them [9]. Since then, we have been using DFA, in our study. Finally, we found that naturally dying crab's heart exhibits a low scaling exponent (about 0.7), and crabs underwent an unpredictable death, which exhibited a high exponent, spanned 1.2–1.5 [10].

We found that natural-death crabs experience a hyperkalemia. Biology can explain the mechanism. Cell death leads to puncture of the cells. The more cells die, the more potassium leaks into the circulation, where the concentration of potassium ions is $\sim$27 times lower, than inside the cell. This potassium leakage from the dead cells causes depolarization of myocardial cell membrane. Depolarization increases the rate of discharge of pace-making heart muscle cells. The outcome of the chain reactions were detectable as a high rate of heartbeats. It is well known that a human, who is near an end exhibits a high heart rate over 200 beat per min (BPM). In our study, sea lice crustaceans, *Ligia exotica*, showed a high heart rate, over 300 BPM, when they died. The largest species of dragonflies, a native of Japan, *Anotogaster sieboldii*, also showed a high heart rate, over 250 BPM. Natural death proceeded gradually, resulting from a gradually increasing number of dead cells. Therefore, from the heart rate, one can notice that the subject is dying, so one can predict a near future event in case of natural death.

Surprisingly, in crab-heart experiments, we encountered an abnormal death that was different from the above-mentioned predictable death; it was an unpredictable death at a high exponent. We noticed that a blood condition of unpredictable-death crabs was normal, because the heart rate did not increase until death. However, interestingly, we noticed that myocardial cells were partially injured by the penetration of EKG electrodes. We conceived the reason why sudden death occurred. In general, sudden death occurs while body cells are normal and heart muscle cells are partially damaged. In such a condition, the pump (heart) was not able to cope with the oxygen demand of body cells, including myocardial cells. The pump gave up working, especially when the acceleratory nerves commanded extraordinarily increased work. That is the heart attack: i.e., unpredictable-death. This unpredictable-death of model animals was comparable to the human ischemic hearts' event. One can recall a sudden death, such as professional athletes. Through the experiments on invertebrate model animals, we learned and found that the scaling exponents are reliable parameters.

Exponentiation is a mathematical operation, written as n^{α}, involving two numbers, the base n and the exponent (or power) α. In our study, the base is a box-size of a heartbeat. DFA calculates α, which is the scaling exponent. Theoretically, n is infinite. But it is impossible to record an infinite length of EKGs. Technologically, how long must we record an EKG for the practical use of DFA in medicine? Which size of box in DFA (see [2, 4]) is required? The answers were not given previously, especially for the field in biology and medicine, instead of in the field of nonlinear dynamic theory. We needed to solve the problems practically.

Dynamic systems are systems that change over time and that can autonomously generate complexity and form. The current state is a function of pervious states and in turn is the basis for future states. In biology, fractal and scaling exist everywhere [11] (Figs. 47.2 and 47.3). The figures show examples of scale-invariance in biology. One can see that this plant's fractal is the results of plant-cells' development over time (Fig. 47.3a and b). However, this scaling does not continue to infinity (see Fig. 47.3c). Biological morphogenesis does not show an infinite feature. There is a limit in biological scale-invariance. When we use DFA in biology, box length (box size in DFA) is limited. In tree structure, the size is confined from 1 mm to 10 m, from leaf, branch, and to trunk (Fig. 47.3). Thereby, in case of a tree, a range of the values of n was confined to [1; 10,000] in mm.

When conducting the DFA on the heartbeat data, at first, we did not know the value of n. We have investigated some hundreds of hearts by our DFA program

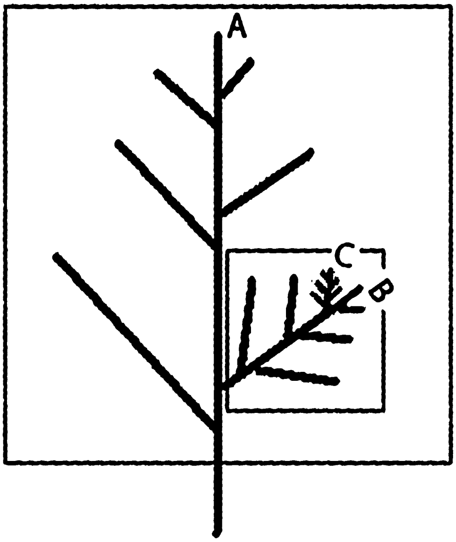


Fig. 47.2 Diagrammatical representation of fractal and scale-invariance in biology. *A*, *B*, and *C* resemble in each structure, different scale from leaf to trunk in size

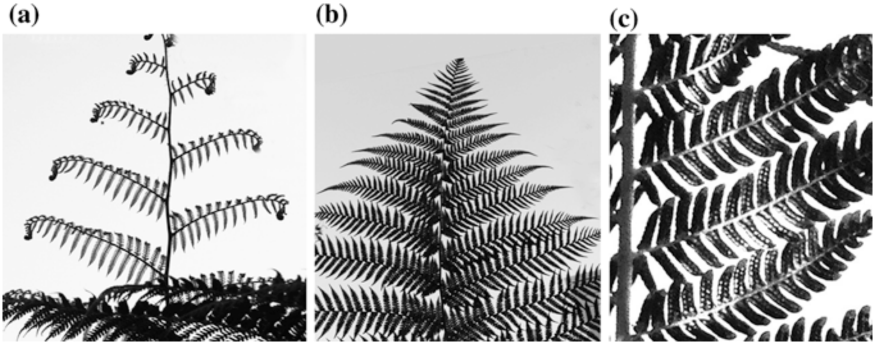


Fig. 47.3 Fern tree leaf pictures. *Photo shows* **a**, scale-invariance structure develops over time. **b** Developed structure resembles with those shown in Fig. 47.2. **c** Fractal disconnects at the veins of a leaf. *Photo* taken by an author at near the University drive, Berkeley, CA, USA

and already found out that a proper n range was confined to [30; 270] [12]. As long as we use our DFA program, DFA computation with this length of heartbeat-numbers, n, guaranteed a good estimation of scaling in heartbeat analysis. The period length from 30 beats to 270 beats roughly corresponds to the length of recordings of EKG from 0.5 to 3 min, respectively. This period of time indicates that it is the period for human keeping the memory in a stationary state. It is only during a restricted time-period, for 3 min.

47.3.2 Estimation Accuracy

DFA is the idea of dividing the accumulated or integrated series into boxes of equal length, n, and to fit a regression line of each box to represent a local trend. This trend is then subtracted from the integrated time series. DFA calculates the corresponding fluctuations, $F(n)$ (see [4] in details). This computation is repeated over all box sizes. A linear relationship between log $F(n)$ and log n indicates the presence of a power law scaling $F(n) \propto n^{\alpha}$, thus fractality. The slope of the regression line relating log $F(n)$ to log n estimates the scaling exponent α.

DFA calculates the positive slope of the line relating to log $F(n)$ and log n. DFA calculates how much variance ($F(n)$) is accounted for by each box-size (n) (heartbeat number). However, it is essential to know whether or not our DFA program is accurately reflecting to the state of a real world data, because estimation accuracy depends on the order of transformation steps [4]. We therefore compared results of our computation program with those of "original idea" that is Peng's program [2]. We confirmed accuracy of our DFA (Figs. 47.4 and 47.5). The two computations showed almost identical results.

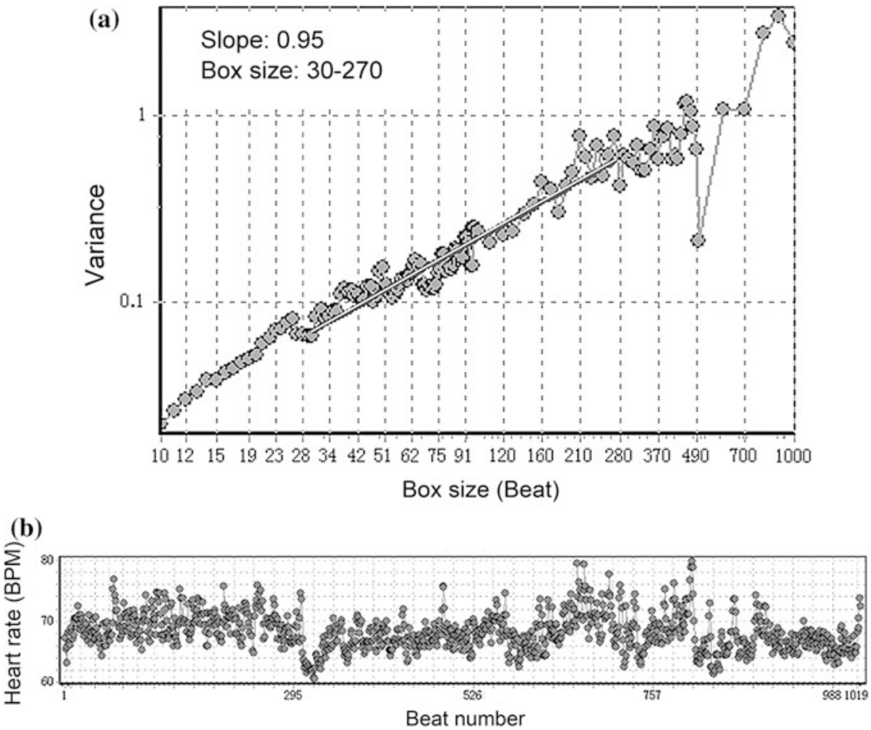


Fig. 47.4 Results of our DFA. Female age 60s. **a** The scaling exponent (α) was 0.95 with box size 30–270. **b** Heartbeat-interval time series upon which DFA applied

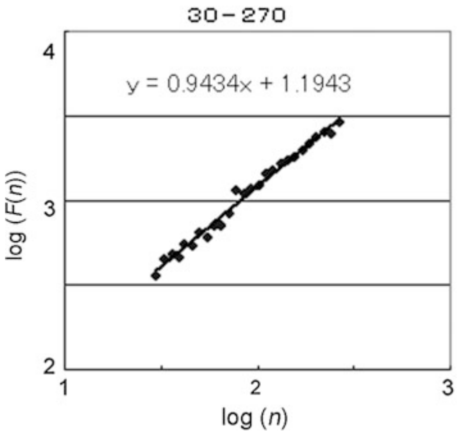


Fig. 47.5 Results of Peng's DFA (see [2]) applied on the same data shown in Fig. 47.4. Female age 60s. The scaling exponent, (α) was 0.9434 with box size 30–270. *n*: box size

47.3.3 Scaling in Human Heartbeat

Figures 47.6 and 47.7 show the results of DFA for three persons. EKGs were simultaneously recorded in a room sitting together side by side, with talking and laughing, for about 40 min. Subject-A exhibited a normal healthy value, 1.04. His heart was perfectly normal in terms of DFA. As for the subject-B, on first sight of time series (Fig. 47.6b), we could not find any significant symptoms. However, his α was 0.85, which was lower than normal value. He mentioned that he feels PVCs especially at mid night (no PVCs in Fig. 47.6b). Years ago, he was admitted to the hospital to checkup although no significant problem was found. We considered that he had not a perfect health condition in terms of DFA. As for subject-C, time series exhibits apparent PVCs (see asterisks, *). His value was very low, 0.72. He mentioned that the number of occurrence of PVCs sometimes increase up to 6 times per min. Although this value is a benign value according to a medical doctors' guideline (exceeding 10 times per min is border line), our DFA seemed to be detecting a hidden abnormality in his system though we did not identify it.

We have so far examined over 300 subjects (not in the hospital) aged 5–88. More than half subjects exhibited unhealthy scaling exponent (never near 1.0 in box size range [30; 270]). Subject-B and subject-C were representative volunteers whom we met. Ironically, subject-A was a healthy but atypical example. Detailed large cohort investigations are required to gather statistics, but we believe that the concept of tailored medicine and healthcare by DFA could be reliable and more helpful than statistics. If the DFA reveals that one has the standard exponent of 1.0, one can never be at a loss. We met a Russian friend researcher (age mid 30s), who have had a valve operation (mitral valveloplasty) a year before. Our DFA revealed that he had the exponent of 1.0, himself and his wife was very relieved. Despite these good results his doctor already told him that the operation was very successful, he and his wife told us, they were happy to get a double confirmation.

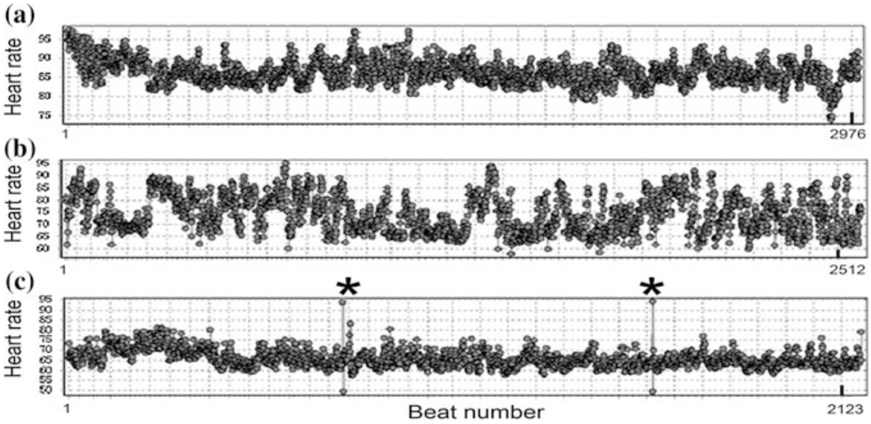


Fig. 47.6 Heartbeat-interval time series, simultaneously recorded from three male professors at the medical school office. Age: early 50s for **a**, 65 for **b**, and 61 for **c**. Subject **c** exhibited two premature ventricular contractions (*asterisks*)

47.3.4 Quantification of Stress

Stress is a physiological reaction of an organism to an uncomfortable or unfamiliar physical or psychological stimulus. The stimuli induce biological changes as the results from activation of the sympathetic nervous system, including a heightened state of alertness, increased heart rate, and so forth. We can define stress in this manner. However, we are not able to quantify stress efficiently. In fact, we can hardly determine if an organism is experiencing stress in response to the stimuli.

We have found that the Japanese spiny lobster, (15–25 cm in size) while in a relaxed condition in a shelter, exhibit, on/off switching patterns of heartbeat sequences; i.e., alternating heart rates from a high rate of 50–70 BPM to extremely low rate of 5–15 BPM [13] (see Fig. 47.8a). However, during stressful states, the lobster does not exhibit the alternating pattern of a heartbeat, but exhibits the continuous beating pattern of 70 BPM (see Fig. 47.8b). The continuous pattern lasts quite a while, as long as the stress stimuli exist. The continuous pattern is the physiological consequence of discharge of cardio-regulatory nerves, i.e., an increased cardio-accelerator discharge at about 60 Hz, and simultaneously occurring cessation of the inhibitory nerve discharge [13].

We therefore focused attention on the difference of pattern of heartbeats between relaxed and stressful states, and challenged to quantify stress by DFA. Figure 47.8 shows the pattern of heartbeats of relaxed lobsters and stressful lobsters. We measured heartbeat intervals of EKG data and constructed a time series of heartbeat-intervals. Figure 47.9 shows a part of time series (578 beats) corresponding to both, relaxed and stressful states. Then we conducted DFA and found that relaxed lobsters in shelter, exhibited a normal scaling exponent of 1. And stressful lobsters being handled by humans, exhibited a lower scaling exponent of 0.6 (Fig. 47.10).

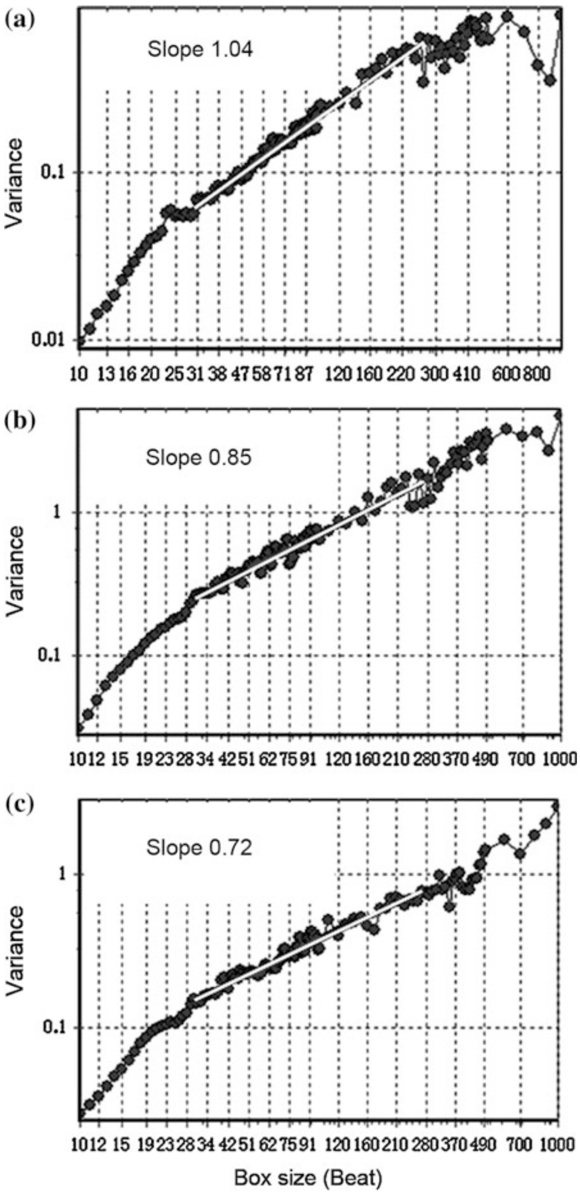


Fig. 47.7 Results of DFA on data shown in Fig. 47.6. Inset, estimated scaling exponents, calculated the slope of the line. Box size 30–270

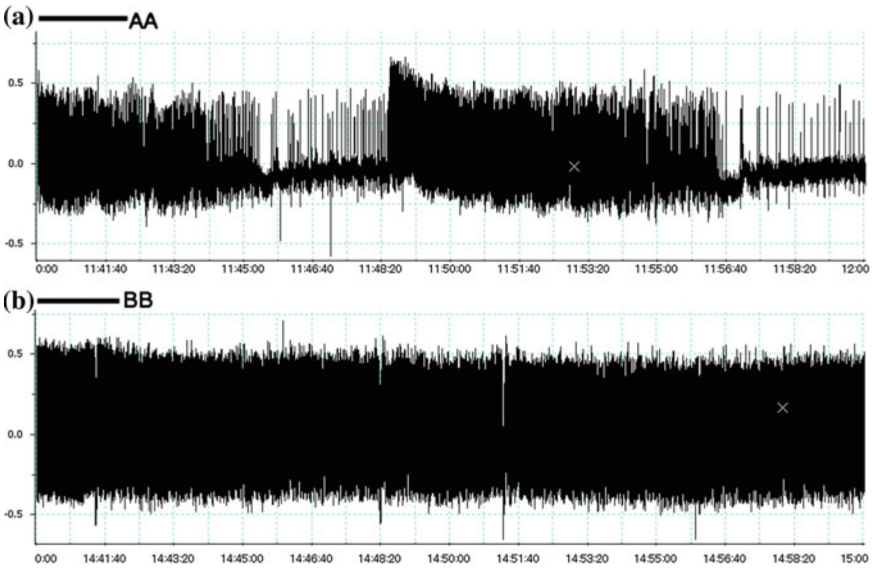


Fig. 47.8 EKGs of Japanese spiny lobster, *Panulirus japonicus*, for 20 min. **a** Lobster was at rest in a shelter under the sea water tank. Lobsters' heart at rest exhibits alternating on/off pattern. **b** This lobster was receiving significant stress under the condition of the micro-dialysis blood sampling experiment

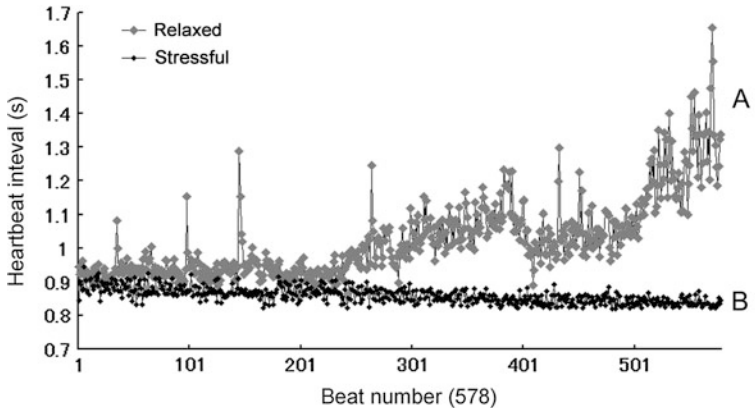


Fig. 47.9 Interval time series calculated from Fig. 47.8. *A*, Relaxed lobster. *B*, Stressful lobster. *A* and *B* correspond to *AA* and *BB* in Fig. 47.8. Only 578 beats shown

47.3.5 Non-Biological Use

Abovementioned results gave us a notion: DFA can be applicable to any cyclic movement, if we observe it's fluctuation characteristics. We then investigated a break down behavior of an electric motor (Fig. 47.11). We used a motor, which

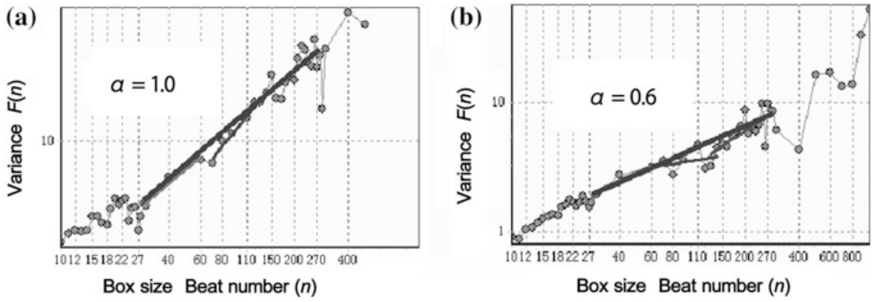


Fig. 47.10 DFA profile. The same lobster shown in Figs. 47.8 and 47.9

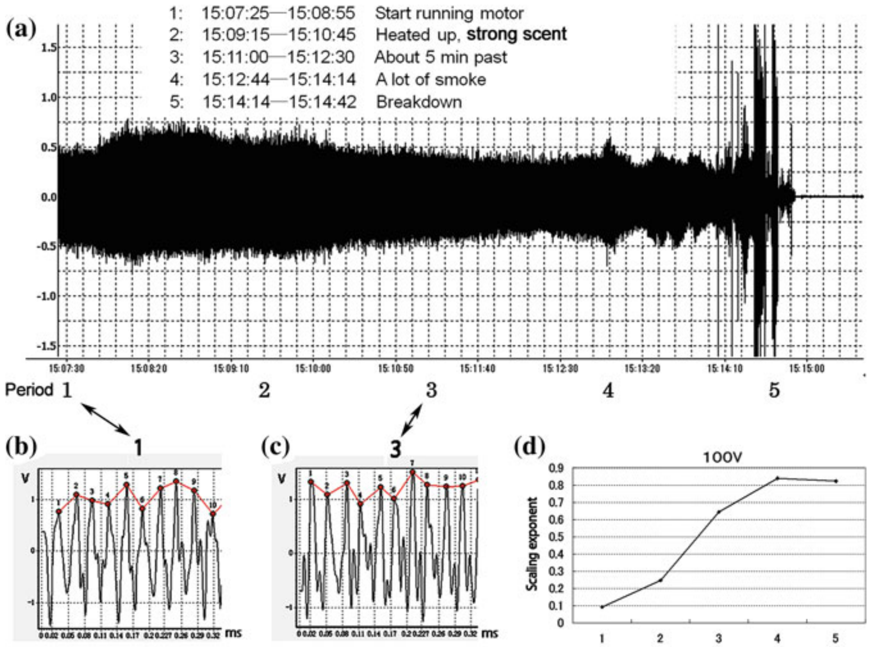


Fig. 47.11 Motor breakdown analysis

was designed for a hand drier, what we see in washrooms of airports and restaurants for example. We covered a motor with glass wool, in order to speed up a breakdown, by heating it up (see Fig. 47.12). A vibratory motion was recorded by a piezoelectric device, fixed on the base. Signals from the piezoelectric sensor transferred to a data logger (Power Lab system) and PC.

Figure 47.11a shows raw data of a vibratory motion of the motor. Figure 47.11a shows the entire period of the experiment; from normal motor running (period 1) to a breakdown condition (period 5). Only 2 min after the start of the running of motor, the motor got overheated and became smoky, with a very strong scent

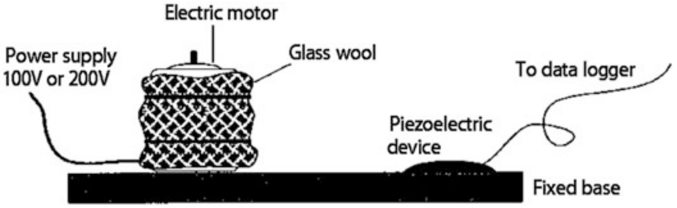


Fig. 47.12 Set up of motor breakdown experiment

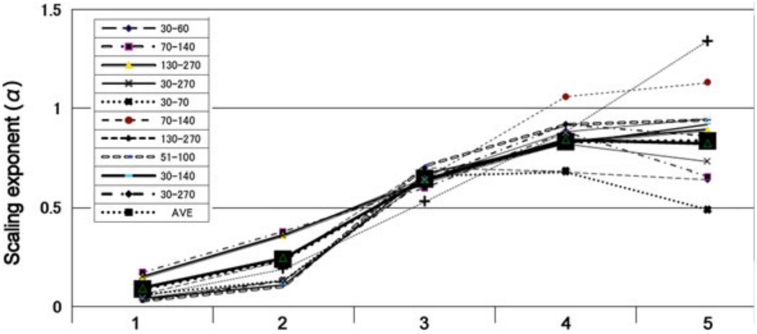


Fig. 47.13 Detailed DFA profile of motor breakdown. The same motor for Fig. 47.12

(period 2). Then, the smoking became very serious (period 3) an irregularly spinning noise appeared (period 3 and 4). Finally, the motor suddenly stopped (period 5, the recording stopped immediately after the breakdown and we escaped from the room: The experimental laboratory was filled with smoke; no fire was ignited. Inset (see Fig. 47.11a) describes what had happened during this period 1–5, regarding to the motor's disorder.

Figure 47.11b and c show a real data of vibration of the fixed base. Sinusoidal patterns are to be seen. One can see that a sinusoidal vibration wave is distorted, even when the motor was running in a normal state (period 1). The vibration wave became noisy over time (period 2, data not shown). Peak intervals were measured (see Fig. 47.11b and c, period 1 and period 3) to construct a peak-interval time series. We applied a DFA technique to this cyclic movement of the motor. The scaling exponents were calculated at each period; from period 1–5 (see Fig. 47.11d). One can see that a "healthy" motor exhibits an extremely low scaling exponent 0.1 (anti-correlation in terms of mathematics and physics). We found out that when a malfunction of the motor develops, the scaling exponent increases (Fig. 47.11d). Detailed transitional changes of the scaling exponent is shown in Fig. 47.13. The scaling exponents were calculated at a various "Box size". We can conclude that any motor, working at the range of scaling exponents over 0.5, (white noise fluctuation) must be checked by engineers at maintenance, to protect against future accidents.

Not only heartbeat fluctuations, but also motor vibrations can be analyzed by DFA. The scaling exponent is quite useful. A healthy heart exhibits 1.0. Healthy man-made structure exhibits less than 0.5. We consider that the collapsing bridge of the highway (Minneapolis, Minnesota, USA, 2007) and the collapsing tunnel of the highway (Yamanashi, Japan, 2012), should be checked before collapsing, by using DFA.

47.4 Discussion

Many people are introduced to the visual world of nonlinear dynamics through a never-ending stream of fractal patterns cascading towards them from deep within their computer screens [14]. The virtual space, generated by computers, seems to be an ideal environment for exhibiting their stunning properties [14].

Unlike computer screens, empirical data in nature, such as fractals in tree-structure and in the heartbeat, is not generated in a never-ending ideal manner. Fractal patterns are found in limited space, indeed a tree fractality range was confined to [1; 10,000] in mm and heartbeat fluctuation fractality was confined to [30; 270] in beat numbers. We showed that DFA works under those limited environment, not under an infinite environment. Despite not infinite, using DFA, we discriminated a healthy heart, unhealthy heart, dying heart, and stressful heart. Stress, particularly its profound, long-lasting effects on behavior and health is a significant health concerns in our days. In the present article, we showed that stress is measurable by our DFA technology. The heart is an opening of mind.

It was in the 80s–90s when Goldberger, Amaral, Hausdorff, Ivanov, Peng, Stanley and colleagues have emphasized the potential utility of DFA in life sciences [6]. Numerous empirical studies identified a noise in human behavior, including noise in heartbeat behavior [4]. However, practical medical use of DFA technology has not progressed to a great extent. Our temporary guideline for determining the wellness of the heart by the scaling exponent, is that a value near 1.0 (specifically, 0.90–1.19) is healthy.

The fluctuation analysis (i.e., DFA) was a potential helpful tool in medicine for the early identification or physiological disorders, as it reveals information that is not provided by an EKG. Unlike HRV [15] (i.e., heart rate variability, the power spectrum, PSD), the excelling point for DFA, is that it has a base line value persistence of one (1), like a standard body temperature (37 °C), a standard blood pH (7.4), and so on. DFA is simple as a tool that everyone could use. No two individuals are ever the same in terms of molecular biology, thus supporting the concept of providing individually tailored medicine and healthcare.

47.5 Conclusion

The scaling exponents could determine whether the subjects are under sick or healthy conditions on the basis of cardiovascular neurophysiology. DFA is practically a useful, numerical method for quantifying the degree of wellness and the transition from sickness to wellness and vice versa. DFA is a simple tool, such as a clinical thermometer and a blood-pressure gauge. Our temporary guideline for determining the wellness of the heart by the scaling exponent is, a value near 1.0 (specifically, 0.90–1.19) is healthy. The present work was partially appeared as the congress proceedings [16].

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 23500524 to TY. We thank financial support of No. 4DQ404 from DVX Inc. Tokyo, Japan, and No. 22DG405 from NOMS, Co. Ltd., Nagoya, Japan. We express our gratitude to T. Tsuruta at Soft Club Co. Ltd. Tokyo, Japan for decades-long understanding and support of our Neurobiology research. We thank G. Witte for correcting our poor English.

References

1. Roger VL, Go AS, Lloyd-Jones DM et al (2012) Heart disease and stroke statistics—2012 update: a report from the American Heart Association. Circulation 125:e2–e220
2. Peng C-K, Havlin S, Stanley HE, Goldberger AL (1995) Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series. Chaos 5:82–87
3. Stanley HE, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M (1993) Long-range power-law correlations in condensed matter physics and biophysics. Phys A 200:4–24
4. Stadnitsuki T (2012) Some critical aspects of fractality research. Nonlinear Dyn Psychol Life Sci 16:137–158
5. Stanley HE (1995) Phase transitions. Power laws and universality. Nature 378:554
6. Goldberger AL, Amaral LAN, Hausdorff JM, Ivanov PC, Peng C-K, Stanley HE (2002) Fractal dynamics in physiology: alterations with disease and aging. PNAS 99(Suppl 1):2466–2472
7. Katsuyama T, Yazawa T, Kiyono K, Tanaka K, Otokawa M (2003) Scaling analysis of heart-interval fluctuation in the in situ and in vivo heart of spiny lobster, Panulirus japonicus. Bull Housei Univ Tama 18:97–108 (in Japanese)
8. Kobayashi M, Musha T (1982) 1/f fluctuation of heartbeat period. IEEE Trans Biomed Eng 29:456–457
9. Yazawa T, Kiyono K, Tanaka K, Katsuyama T (2004) Neurodynamical control systems of the heart of Japanese spiny lobster, Panulirus japonicus. IzvestiyaVUZ Appl Nonlinear Dyn 12(1–2):114–121
10. Yazawa T, Tanaka K, Katsuyama T (2007) DFA on cardiac rhythm: fluctuation of the heartbeat interval contain useful information for the risk of mortality in both, animal models and humans. J Systemics Cybern Inform 5(1):44–49
11. Barnsley M (1988) Fractals everywhere. Academic press, San Diego
12. Yazawa T, Tanaka K, Kato A, Nagaoka T, Katsuyama T (2007) Alternans lowers the scaling exponent of heartbeat fluctuation dynamics in animal models and humans. In: WCECS2007 proceedings, vol 1. San Francisco, pp 1–6
13. Yazawa T, Katsuyama T (2001) Spontaneous and repetitive cardiac slowdown in the freely moving spiny lobster, Panulirus japnicus. J Comp Physiol A 187:817–824

14. Taylor RP (2012) The transience of virtual fractals. Nonlinear Dyn Psychol Life Sci 16:91–96
15. Stauss HM (2003) Heart rate variability. Am J Physiol Regul Integr Comp Physiol 285:R927–R931
16. Yazawa T, Shimoda Y (2012) DFA applied to the neural-regulation of the heart. In: Lecture notes in engineering and computer science: proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, pp 715–720, 24–26 Oct 2012

Chapter 48
Identification of Diabetic Neuropathic Conditions of Simulated Ulnar Nerve Response Using Prony's Method

V. Sajith, A. Sukeshkumar and Jinto Jacob

Abstract Using a portable biopotential amplifier (Electroneurograph), the evoked response of the Ulnar nerve has been recorded from patients with moderate to acute diabetic neuropathic conditions. The ENG signals are again simulated and the features of these signals are extracted using Prony's feature extraction method. The poles and resonance frequencies of ENG signals are used as features. A multilayer feed forward neural network classifier is applied to these features for the classification of Neuropathic Conditions. Identification of moderate to acute diabetic neuropathic conditions at various stages is the primary significance of this technique which helps the doctor for better diagnosis and to go for advanced treatment.

Keywords Artificial neural network · Diabetic neuropathy · Electroneurograph · Nerve stimulation · Prony's method · Ulnar nerve

48.1 Introduction

Diabetic neuropathy is the most common and troublesome complication of diabetes mellitus leading to great morbidity and resulting in a huge economic burden for diabetes care. Diabetic patients have a greater risk factor of developing nervous

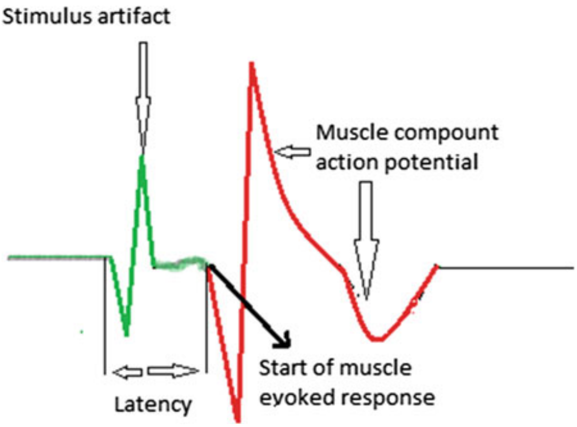
V. Sajith (✉) · A. Sukeshkumar
University of Kerala, Kerala, India
e-mail: sajithvijayaraghavan@gmail.com

A. Sukeshkumar
e-mail: sukeshkumarcet@yahoo.com

J. Jacob
BSNL, Kerala, India
e-mail: jinto83@gmail.com

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_48,

Fig. 48.1 Typical normal ulnar nerve


Stimulus artifact
Muscle compount action potential
Latency
Start of muscle eyoked response

disorders when compared with non-diabetic subjects. The presence of other vascular complications such as peripheral vascular disease in diabetes increases the risk of diabetic foot complications. However, the progression of neuropathy can be reduced by early detection and intervention [1].

In nerve conduction studies, primarily nerve conduction velocities are considered as one of the most sensitive indices of the severity of diabetic neuropathy. Electroneurography is the recording and study of action potential propagation along peripheral nerves. Neurological examination of peripheral nerve function refers to the nerve conduction study which is based on evoked potential. Evoked potentials are obtained when electric pulses are passed through the excitable tissues like nerves and muscles. For nerve conduction studies the parameters of the evoked responses required are the amplitude, latency and area of the responses recorded (Fig. 48.1).

With this method it is very easy to conduct study of neurological examinations like conduction velocity measurement. An economic, versatile, compact and portable biopotential amplifier has been designed and implemented for measuring the latency to calculate the conduction velocity of ulnar and median nerves of patients with diabetic neuropathic conditions [2]. The specific application is developed according to the requirements of the experimental research practice on the study of compound action potentials of nerves in the field of neurophysiology [3]. Early detection of conduction failure of ulnar nerve in patients with diabetic type 1 and 2 is possible using conduction velocity measurement [4]. For the identification of various neuropathic conditions a mathematical model of the evoked response of diabetic patients using discrete state space modeling is done [5]. This paper establishes the identification of moderate to acute diabetic neuropathic conditions of simulated ulnar nerve response using prony's method.

Many methods and algorithms from signal processing theory have been proposed, described and implemented over the past few years to extract feature from signals such as total least squares based Prony's modeling algorithm [6], autoregressive model [7], wavelet transform [8], multiple signal classification (MUSIC) algorithm [9]. Drawback of these techniques mentioned above is the significant amounts of computation and processing time for extracting the features and the classifications technique employed.

The objective of the this work is to apply Prony's method to represent the ENG signal as a sum of exponentials in order to classify certain ENG identification based on the calculation of the resonance frequencies and the accompanied poles. Resonance complex frequencies of a signal can be used as useful feature for signal discrimination and identification.

The advantage of Prony's method is its simplicity and applicability in real time where a big amount of data may exist. The resonance frequencies computed were classified using multilayer feed forward network classifier models to determine the different classifications of diabetics.

The article is organized as follows. Section 48.2 introduces the method proposed. Section 48.3 describes the preprocessing procedure used in this work. Section 48.4 gives a review on the feature extraction method of the proposed work. Section 48.5 gives the classifier model used. Section 48.6 discusses implementation procedure. Section 48.7 discusses simulation results.

48.2 Methodology

48.2.1 A. Normal Method

In order to study the function of neurons it is usually the practice to stimulate the nerve and study the action potentials generated by it. The nerve can be stimulated to greater activity by either electrical or chemical stimulation. Here we are interested only in electrical stimulation which is easily controllable. The electrical stimulation can be either constant voltage or constant current. It actually involves passing an appropriate electric current through a region of tissue and thereby causing the neurons in the area to be depolarized.

The nerve stimulation techniques are performed with a variable stimulus generator. The stimulus frequency, stimulus intensity and stimulus duration must be known, adjustable and have an ON/OFF control.

The input portion of the instrumentation amplifier which is named head stage is designed separately and the inputs from the electrodes are given to the head stage by means of shielded coaxial cable. Also the electroneurograph is properly shielded and grounded to minimize the noise effect.

Offset control is provided for base line shift. The stimulating and recording electrodes were placed in appropriate positions and evoked responses were recorded. The block diagram of the recording setup is shown in Fig. 48.2.

The normal method adopted for the measurement of conduction velocity is by stimulating at two different points along the nerve and measuring the latency for

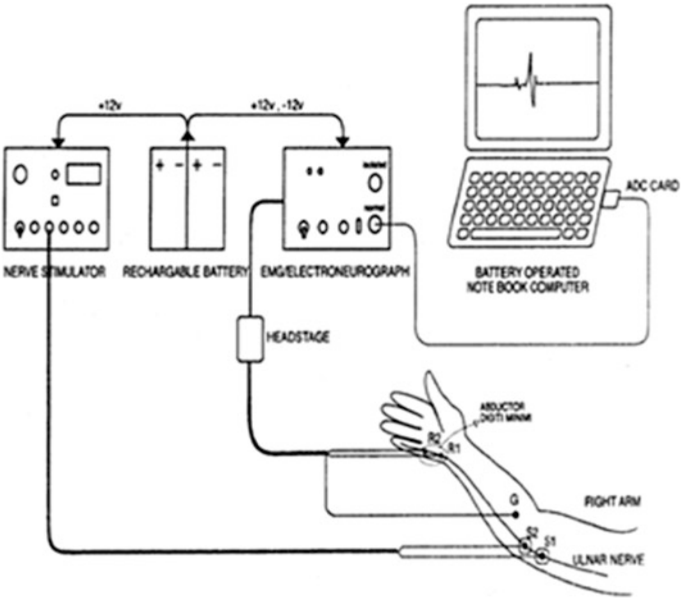


Fig. 48.2 Recordical setup

each electrical response recorded from the muscle. For each stimulation, the time between the start of the stimulus and the onset of the response is measured which is termed as the ‘latency’. The length of nerve segment is obtained by measuring on the surface of the skin the distance between the cathodes when placed for each stimulation. The conduction velocity is obtained by dividing the nerve length between the two stimulating points by difference in latency which is usually expressed in meters per second. It has been observed that a mean conduction velocity of 55.71 m/sec and a standard deviation of 3.01 m/sec was measured for the normal Ulnar nerve [3]. The electrode positions for the recordings and responses of the Ulnar nerve and are shown in Fig. 48.2.

Figure 48.3 shows the Ulnar nerve response of a normal patient. Figure 48.4 depicts the Ulnar nerve response of a moderate diabetic patient and Fig. 48.5 shows the Ulnar nerve response of an acute diabetic patient.

48.2.2 B. Proposed Method

The different stages of the proposed methods to identify diabetic neuropathy based on ENG signals are depicted in Fig. 48.6. It mainly consists of four stages: ENG database, Preprocessing stage, Feature extraction stage, Classification stage. The digital ENG signals are taken using above setup and are stored to form database. These signals are first preprocessed to remove baseline wander, power line

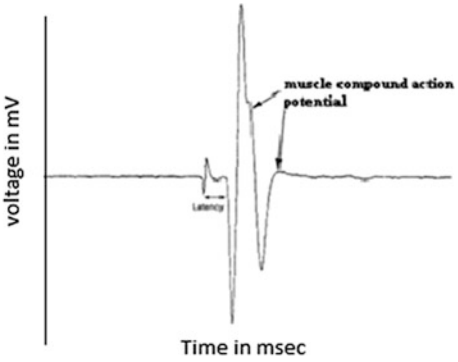


Fig. 48.3 Nerve action potential-normal

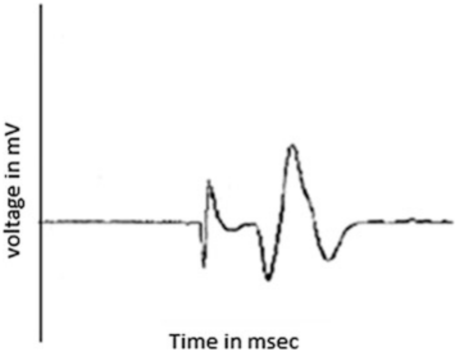


Fig. 48.4 Nerve action potential-moderate

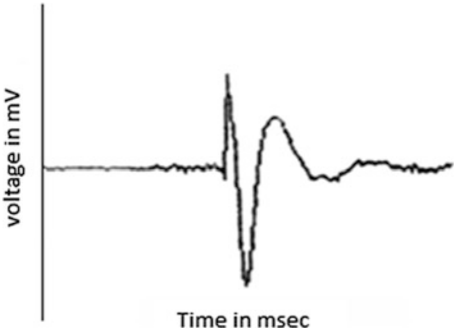


Fig. 48.5 Nerve action potential-acute

interference and high frequency noise. Now Prony's method is applied to extract feature from ENG signals [10]. Finally, neural network classifier models are used to test those features and the diagnosis is made.

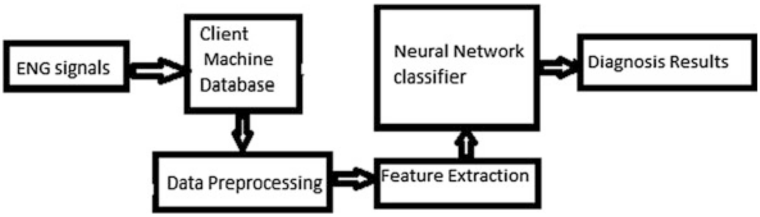


Fig. 48.6 Proposed system

48.3 Preprocessing Stage

All ENG data's have been filtered to remove the noise that may influence the signal including, baseline wander, artifact, and power line interference. The presence of these noise sources in the signal may mislead the feature extraction and classification.

Butterworth high pass filter is designed to remove these low bands of frequencies. The cutoff frequencies of the high pass filter is selected as 150 Hz.

48.4 Review on Prony's Method

Prony's method [10] is a technique for modeling sampled data as a linear combination of damped exponentials as given below

$$\hat{x}(t_n) = \sum\nolimits_{k=1}^{M} h_k z_k^{(n-1)} \tag{1}$$

where n = 1,2,3,…N (N-samples),

$$h_k = A_k e^{j\Phi_k},$$

$$z_k = e(\alpha_k + j\omega_k)T_s,$$

T_s is the sampling period, A_k is the amplitude, α_k is the damping factor, $\acute{\omega}_k$ is the angular velocity, Φ_k is the initial Phase and k is the exponential code.

The Prony's method evaluates the model parameters by solving two sets of linear equations. To find the first set of linear equations, let us consider the polynomial having the z_k as its roots:

$$F(z) = \prod\nolimits_{k=1}^{M} (z - z_k) = \sum\nolimits_{M=0}^{M} \mathrm{a(m)} z^{M-m} \tag{2}$$

with a(0) = 1.

Using Eq. (2) and with proper manipulations on relations(1), this results as:

$$\sum\nolimits_{M=0}^{M} a(m)x(n-m) = \sum\nolimits_{M=0}^{M} a(m) \sum\nolimits_{k=1}^{M} h_k z_k^{(n-m-1)} \tag{3}$$

with n = M + 1,M + 2,…N.

In (3), using the position $z_k^{(n-m)} = z_k^{(n-M)} z_k^{(M)}$, the right-hand side contains the polynomial F(z) evaluated at its root z_k, then:

$$\sum\nolimits_{k=1}^{M} h_k z_k^{(n-m)} \left\{ \sum\nolimits_{m=0}^{M} a(m) z_k^{(M-m-1)} \right\} = 0 \tag{4}$$

From (3), using (4), this gives:

$$\sum\nolimits_{m=0}^{M} a(m)x(n-m) = 0 \tag{5}$$

with n = M + 1,M + 2,. . . ,N. The (N-M) relations (5) constitute a linear equation system in M unknowns, i.e. the a(m) coefficients.

Using N = 2 M samples the system (5) represents an M equation system with the same number of unknowns. For practical situations there exists a non-trivial solution due to Nyquist-Shannon digital sampling theorem, the number of data points N usually exceeds the minimum number needed to fit a model of exponentials, i.e. N > 2M. In this over determined data case, the linear equation (5) must be modified to:

$$\sum\nolimits_{m=0}^{M} a(m)x(n-m) = e(n) \tag{6}$$

The estimation problem bases on the minimization of the total squared error, e(n):

$$E = \sum\nolimits_{n=m+1}^{N} |e(n)|^2 \tag{7}$$

With the knowledge of the a(m) coefficients, the roots z_k of the characteristic polynomial (2) can be calculated. The damping factors and frequencies of each component may be evaluated from these known roots. Finally to calculate the amplitudes and phases of each component it is necessary to solve the second set of linear equations (1) in the unknown's h_k. The poles and the accompanied resonance complex frequencies, f_k of the ENG signal can be directly calculated as, z_k/T_s.

48.5 Classification Stage

There is several classification models like Artificial Neural Network-Nearest Neighbor, Linear Discriminate Analysis and Support Vector Machines are used for the diagnosis. In this work we have used ANN for the classification.

Multilayer feed forward network named the multilayer perceptron (MLPs) is employed as a class of neural networks. Usually, MLP is made up of several layers of neurons. Each layer is fully connected to the next one. MLP consists of two phases, the training phase and the testing phase. During the training phase, the features are applied at the input and the corresponding desired classes are at the output of MLP classifier. A training algorithm is executed to adjust the weights and the bias until the actual output of the MLP matches the desired output and performance satisfaction is reached. In the test phase, a set of test features, which are not part of training features, are applied to the trained MLP classifier to test the classification of the unknown features.

48.6 Implementation Procedure

Initially, signals were filtered using a Butterworth high pass filter with cutoff frequency of 150 Hz to reduce the noise. Figure 48.7, shows a normal signal corrupted with baseline drift and power line interference noise. In Fig. 48.8, shows a noise free signal as a result of applying the Butterworth high pass filter. Similarly Figs. 48.9 and 48.10 corresponds to moderate and acute. Figures 48.11 and 48.12 are their preprocessed waveforms.

Testing and training sets are separately formed. The numbers of testing and training samples are shown in Table 48.1. We randomly select the given number of training and testing evoked responses of ulnar nerve from the selected recordings.

In our experiments, we used the multilayer perceptron(MLP) as the pattern classifiers. The original data set was separately divided into training and testing groups. Two factors that might affect the efficiency of MLP are the number of hidden layer neurons and synapses initial weigh. Experiments were performed to test the effects of the initial weight and the number of hidden layer neurons parameter.

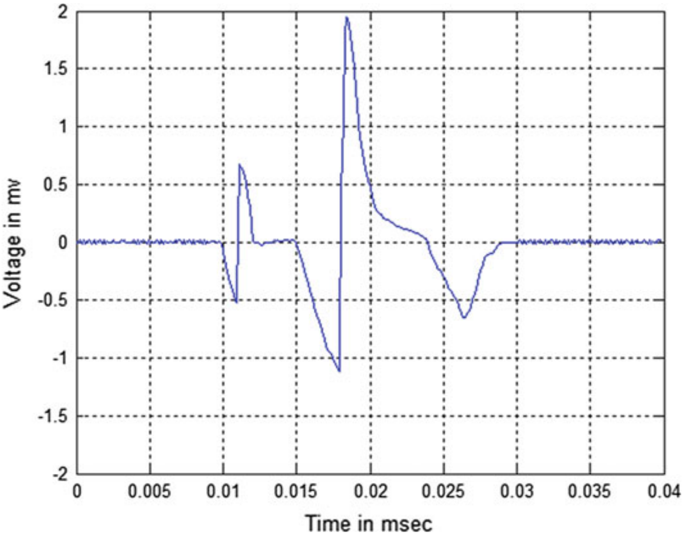


Fig. 48.7 Unfiltered normal signal

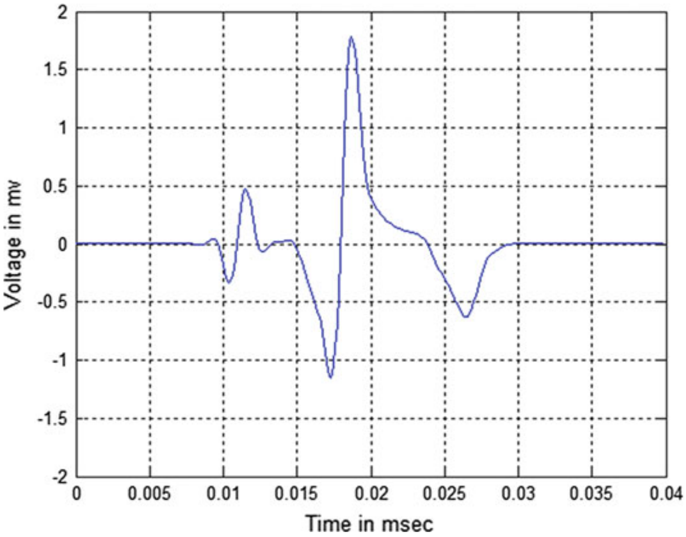


Fig. 48.8 Pre-processed normal signal

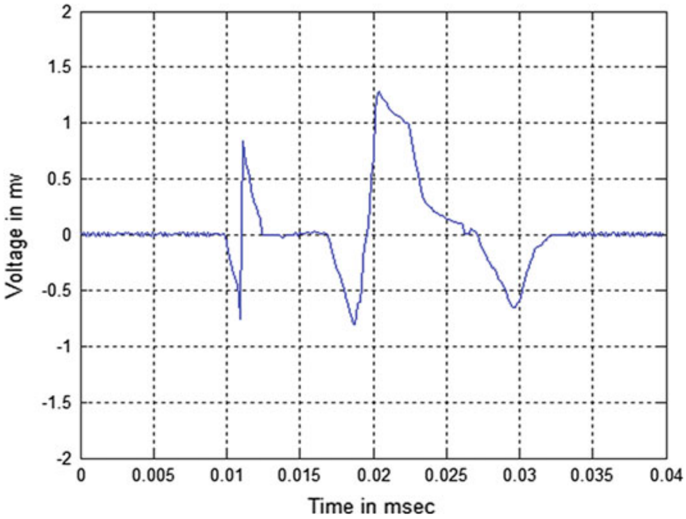


Fig. 48.9 Unfiltered moderate signal

The performances of the classification are evaluated in terms of sensitivity, specificity, and overall accuracy. Sensitivity and specificity are used to evaluate the ability of the classification system to discriminate one class against the other. The sensitivity is calculated as the proportion of positive samples correctly

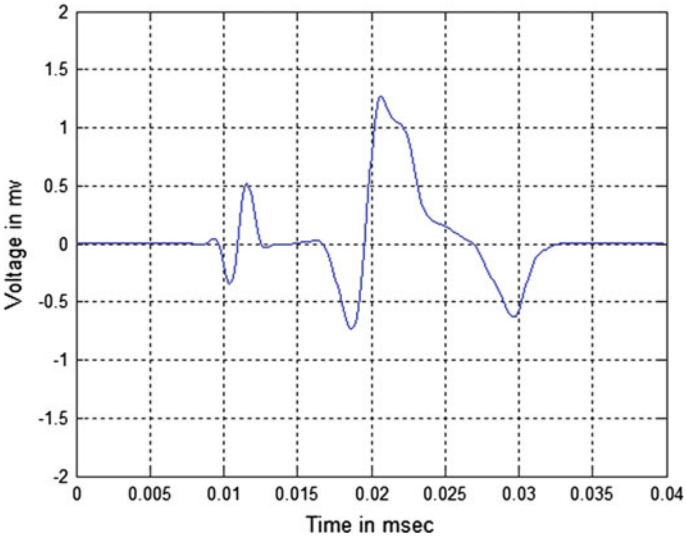


Fig. 48.10 Pre-processed moderate signal

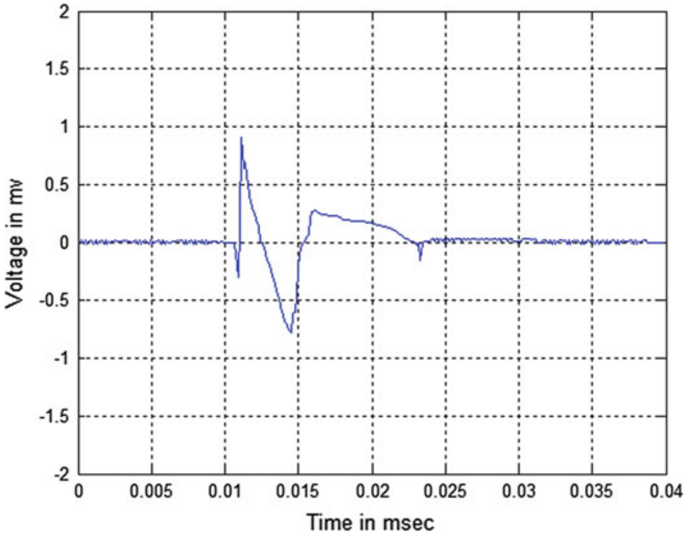


Fig. 48.11 Unfiltered acute signal

assigned to the positive class. The specificity is the proportion of negative samples correctly assigned to the negative class. The overall accuracy is the fraction of the total number of simulated responses correctly classified.

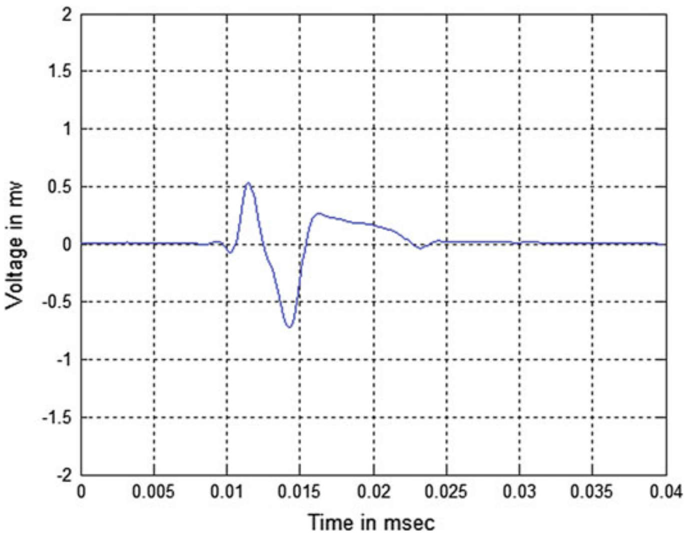


Fig. 48.12 Pre-processed acute signal

Table 48.1 Number of samples used for study

Type of signal	Training samples	Testing samples
Normal	10	50
Moderate	10	50
Acute	10	50

48.7 Experimental Results

The classification results using MLP is given in Table 48.2. The diagonal elements in the table are the number of correctly classified response of specific ENG types using the proposed method [11]. In Table 48.2, MLP neural network generally provides adequate recognition throughout all categories.

Table 48.3 shows the sensitivity, specificity, and overall accuracy of the extracted feature set with MLP classifiers. The results are computed when the number of MLP hidden layer neurons is 55. Average of specificity is 100 % for MLP classifier. Also the average of overall accuracy is 100 %.

Table 48.2 Classification result

Actual/classified	Normal	Moderate	Acute
Normal	50	0	0
Moderate	0	50	0
Acute	0	0	50

Table 48.3 Sensitivity, specificity and overall accuracy

	Type	Average (%)
Sensitivity	Normal	100
	Moderate	100
	Acute	100
Specificity	Normal	100
	Moderate	100
	Acute	100
Overall accuracy		100

48.8 Conclusion and Future Work

The results of this study have significance in diagnosis of diabetic neuropathy at various stages. This also helps the doctor to identify the level of diabetics and take immediate remedial steps.

In future, this work can be extended to any diabetic patients having different stages of diabetic neuropathy.

Acknowledgements The authors take this opportunity to acknowledge generous support and feedback extended by the Principal, College of Engineering, Trivandrum and my colleagues.

References

1. Viswanathan V, Seena R, Nair M, Snehalatha C, Bhoopathy R, Ramachandran A (2004) Nerve conduction abnormalities in different stages of glucose intolerance, neurology India
2. Sajith V, Sukeshkumar A (2008) Design and intra-operative studies of an economic versatile portable biopotential recorder. In: 13th ICBME conference proceedings, Singapore, 3–6 Dec 2008
3. Sajith V, Sukesh Kumar A (2009) Studies on conduction velocity of ulnar and median nerves in patients with moderate to acute diabetic neuropathic conditions, ICOICT 2009. In: International conference on optoelectronics, information and communication technologies, India
4. Sajith V, Sukesh Kumar A (2009) Early detection of conduction failure of ulnar nerve in patients with diabetic type 1 and type 2. World Acad Sci Eng Technol 54:1347–1350
5. Sajith V, Sukesh Kumar A (2009) Mathematical model of the nerve evoked response recorded using electroneurograph in diabetic patients using state space modeling—a new approach. In: International conference on modeling and simulation (MS09), 1–3 Dec 2009, pp 293–296
6. Chen SW (2000) Two-stage discrimination of cardiac arrhythmias using a total least squares-based prony's modeling algorithm. IEEE Trans Biomed Eng 47:1317–1326
7. Ge D, Srinivasan N, Krishnan S (2002) Cardiac arrhythmia classification using autoregressive modeling. Biomed Eng OnLine 1(1):5, 1585–1588
8. Inan OT, Giovangrandi L, Kovacs GTA (2006) Robust neural network based classification of premature ventricular contractions using wavelet transform and timing interval features. IEEE Trans Biomed Eng 53(12):2507–2515

9. Ahmad NNR, Kadkhodamohammadi AR (2008) Cardiac arrhythmias classification method based on MUSIC, morphological descriptors, and neural network, EURASIP. J Adv Signal Process 2(202)
10. Lobos T, Rezmer J, Schegner P (2003) Parameter estimation of distorted signals using prony's method. In: IEEE bologna power tech conference, Bologna/Italien
11. Sajith V, Sukesh Kumar A, Jacob J (2012) Identification of moderate to acute diabetic neuropathic conditions of simulated ulnar nerve response using prony's method, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, USA, 24–26 Oct 2012, pp 1092–1097

Chapter 49
Despeckling of Ultrasound Images of Bone Fracture Using RADWT Based Non-Linear Filtering

Deep Gupta, Radhey Shyam Anand and Barjeev Tyagi

Abstract Despeckling in ultrasound medical images is of great interest. Due to presence of speckle, experts may not be able to extract correct and useful information from the images. This chapter presents a method for despeckling based on a new rational-dilation wavelet transform (RADWT) and non-linear bilateral filter (BLF). The RADWT, a new family of the discrete wavelet transform for which frequency resolution can be varied, provides effective representation of the noisy coefficients. Bilateral filter and thresholding scheme are applied to the noisy RADWT coefficients to improve the denoising efficiency and preserve the edge features effectively. The proposed method also helps to improve the visual quality of bone fracture ultrasound images. The performance of the proposed method is evaluated on the different ultrasound images of bone fracture and results show significant improvement not only in the speckle reduction but also in the edge preservation performance.

Keywords Bilateral filter (BLF) · Bone fracture · Rational-dilation wavelet transform (RADWT) · Speckle · Thresholding · Ultrasound

D. Gupta (✉) · R. S. Anand · B. Tyagi
Department of Electrical Engineering, Indian Institute of Technology Roorkee,
Roorke, India
e-mail: er.deepgupta@gmail.com

R. S. Anand
e-mail: anandfee@iitr.ernet.in

B. Tyagi
e-mail: btyagfee@iitr.ernet.in

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5_49,

49.1 Introduction

Currently, the research in the medical imaging has produced many different imaging modalities for the clinical purpose. The most common imaging modalities used in orthopedics are X-rays, computed tomography (CT), magnetic resonance imaging (MRI) and ultrasound. These medical imaging modalities depend heavily on the computer technology for the creation and display of the medical images [1]. However, all these medical imaging modalities use the ionizing radiation on the basis of different type of the energy sources such as internal, external and combination of both, used in imaging but it may produce important safety concerns for the human beings. Among the medical images from different imaging modalities, Ultrasound B-scan images are widely used. This widespread choice is due to its cost effectiveness, portability, acceptability and safety [1].

In [2], different examinations of 243 apophyseal fractures in adolescents with X-rays and ultrasound imaging are reported, in which 80 cases were diagnosed by X-rays and 97 by ultrasound successfully. In Hübner et al. [3], described the possibility of the diagnosis for the fractures of the children using ultrasound imaging. The authors have also suggested that the ultrasound assessment without radiography should be used in particular cases such as bulge fractures or mildly displaced, simple fractures of the long bones of the fore arm, humerus, femur, lower leg and clavicle. In Rathfelder et al. [4], presented the experimental results of 2006 X-rays which are used to identify the different fractures. In all these experiments, only 345 fractures were diagnosed completely so they described the possible application of the ultrasonography and its good results on the same experiments.

Heiner et al. [5] have demonstrated to be an effective method of rapid identification of long bone fractures. In [6], a simulation model has been prepared for the ultrasound diagnosis for the bone fracture. Evaluation by the ultrasound for fracture detection has the main advantage of immediate clinical correlation and can be made by the people with minimal training in the use of the ultrasound [2, 7, 8]. It is found that ultrasound imaging was more sensitive than conventional radiography for rib fracture diagnosis [9–11].

However, the images obtained from ultrasound imaging are of relatively poor quality. Noise and artifacts are introduced in ultrasound images due to limitation of the acquisition techniques and systems. Among these noise and artifacts, speckle is a main factor which degrades the quality and most importantly texture information present in the ultrasound images. Speckle is considered as a multiplicative noise which has undesirable interference effect on the images. In addition to multiplicative noise, sometimes these ultrasound images also suffer from random additive noise. Due to presence of these noises, diagnosis becomes a time consuming job besides being susceptible to errors depending upon the radiologist experience and expertise. To improve the quality of images, the most important issue is the reduction of these speckles. Sometimes speckle reduction process may suppress the important details/features of the ultrasound images, so speckle reduction algorithms should be designed in such a manner that it smoothens the images without significant loss of information.

Ultrasound speckle reduction methods can be classified in two categories viz. compounding (image averaging) and post-acquisition (image filtering) [12]. Image averaging is usually achieved by averaging a series of uncorrelated ultrasound images in the spatial or frequency domain. These uncorrelated images may be sampled at different times, from different views or with different frequency for same target. It is apparent that such methods suffer from the loss of spatial resolution. Filtering methods are a practical alternative for most clinical applications. It can be classified as single scale spatial filtering such as linear [13], nonlinear adaptive methods [14, 15] etc., multiscale spatial filtering such as diffusion based methods [12, 16–20].

Lately comprehensive efforts have been made to reduce the speckle noise and overcome the drawback of spatial domain filtering, using the wavelet transform. In [21, 22], wavelet thresholding has been proposed as a true signal estimation technique that utilizes the capabilities of wavelet transform for signal denoising. In [23], soft thresholding denoising method is presented in which the main critical task is the selection of the threshold value. Various threshold selection techniques have been reported as VisuShrink, SureShrink and BayesShrink [22, 24, 25]. In [26, 27], the statistical method such as Bayesian approach has been applied in denoising of images and later on, the researchers extended this approach using different noise models for the different noisy wavelet coefficients [24, 28, 29]. The main strength of wavelet thresholding is the capability to handle the different frequency subbands of an image separately. However, the problem experienced is generally smoothening of edges. In [30], wavelet based total variation filtering is also presented in which noisy image undergoes several iterations for suppressing the noise and leads to blurring effect. The bilateral filter was presented as an alternative to wavelet thresholding [31]. It is a non-linear filter and used in spatial domain for edge preserved denoising. In [32, 33], the wavelet transform based bilateral filtering approach has been presented. It provides better denoising and also effectively preserves the edges. This method exploits the potential features of both wavelet thresholding and bilateral filter at the same time. However, wavelet transform based method may introduce many visual artifacts in the denoised images due to fixed wavelet basis and fixed resolution.

Currently research is concentrated in the wavelet domain because of its primary properties like sparsity and decomposition of wavelet coefficients. The sparsity property of the wavelet transform, combined with the capacity for analyzing time frequency information simultaneously within different frequency subbands or temporal resolution, makes the wavelet domain ideal for the problem. Wavelet transform is an effective processing tool for smoothing the signals. However, it has a poor frequency resolution or low Q-factor which limit the effectiveness of this wavelet transform. Selesnick et al. [34] introduced a new family of the wavelet transform, known as rational-dilation wavelet transform, for which the frequency resolution can be varied. This wavelet transform is more flexible and modestly overcomplete, based on rational-dilation. In [35], the authors proposed an enhancement technique using the RADWT based soft/hard thresholding and provide the better results than the existing wavelet based methods. This chapter presents a method which combines the rational-dilation wavelet transform with bilateral filter and thresholding scheme to suppress the speckle as much as possible and retains the edges in ultrasound images of the bone fracture.

49.2 Theoretical Background

49.2.1 Rational-Dilation Wavelet Transform (RADWT)

The RADWT is a discrete wavelet transform with the wavelet variables of time and frequency subband as a position and scale along with the rational-dilation factor. Different types of wavelet transform are used in the different image processing applications such as debluring, sharpening, denoising, compression, interpolation and classification. Wavelet transform are based on the analysis of the signal at different resolution levels. Those applications which require the transforms are invertible; the resolution is doubled from one resolution level to other level. The RADWT is based on the rational-dilation factor between one and two, where the resolution is increased more from one resolution level to another resolution level. It provides more flexibility for varying frequency resolution because the dyadic wavelet transform limits its effectiveness due to low Q-factor i.e. poor frequency resolution in case of oscillatory signal [34] but the RADWT provides a wide range of the constant Q-factor depending upon the dilation factors. The RADWT is implemented using a perfect reconstruction by using two iterated two channel filter bank with rational-dilation factor (p and q) and high pass sampling factor (s). The RADWT is characterized by its Q-factor and its redundancy i.e. oversampling rate [34]. The RADWT is implemented using the iterated filter bank with dilation parameters A = q/p and B = s, which is shown in Fig. 49.1.

In Fig. 49.1, H is a low pass filter and G is a high pass filter in the analysis filter bank which is used for the implementation of the RADWT.

The RADWT is a self-inverting (or form a tight frame) if the analysis and synthesis filter bank shown in Fig. 49.1 have a perfect reconstruction property. If the analysis and synthesis filters are so designed that the output signal is equal to the input signal then the filters are said to satisfy the perfect reconstruction condition. The necessary and sufficient conditions on the filter H and G to insure the perfect reconstruction are given as [34]:

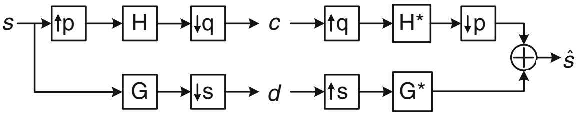


Fig. 49.1 Analysis and synthesis filters for the implementation of RADWT

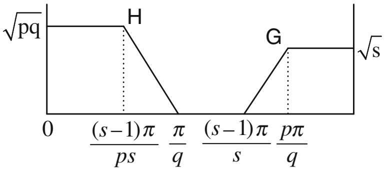


Fig. 49.2 The perfect reconstruction condition for the filter bank [34]

$$H(\omega) = 0 \quad \text{for } |\omega| \in [\pi/q, \pi] \tag{49.1}$$

$$G(\omega) = 0 \quad \text{for } |\omega| \in [0, (1 - 1/s)\pi] \tag{49.2}$$

$$\frac{1}{pq}\left|H\left(\frac{\omega}{p}\right)\right|^2 + \frac{1}{s}|G(\omega)|^2 = 1 \quad \text{for } \omega \in [0, \pi] \tag{49.3}$$

These conditions imply a transition band for the filters H and G with the different widths [34], as shown in Fig. 49.2.

For the perfect reconstruction of the wavelet transform, the length of the signal should be multiple of q and s. The signal length should be multiple of the least common multiple of q and s. Inverse rational-dilation wavelet transform (IRA-DWT) is computed by the transpose of the forward RADWT. For rational-dilation wavelet transform, the wavelet is defined in frequency domain as in [34].

$$\widehat{\psi}(\omega) = \lim_{j\to\infty}\left(\frac{p}{q}\right)^{j/2} G_j\left(\left(\frac{p}{q}\right)^j \omega\right) \tag{49.4}$$

So by choosing the parameter p, q and s, the RADWT can achieve high Q-factor with good time frequency localization. Four stage forward and inverse rational-dilation wavelet transform are illustrated in Fig. 49.3.

49.2.2 Bilateral Filter

Bilateral filter (BLF) is a non-linear filtering process that performs the edge preserved denoising within the spatial domain [31]. The BLF replaces the pixel values by a weighted sum of the pixels in a neighborhood and the weights depend on the spatial distance of the pixel around the neighborhood and the intensity distance around the neighborhood of a pixel. It is achieved by the combination of two Gaussian filters; first one is domain filter while the second is the range filter [33]. So at a pixel location x, the response of the BLF can be computed as

$$\widehat{s}(x) = \frac{1}{C}\sum\nolimits_{y\in N(x)} D_f(x, y)R_f(x, y)s(y) \tag{49.5}$$

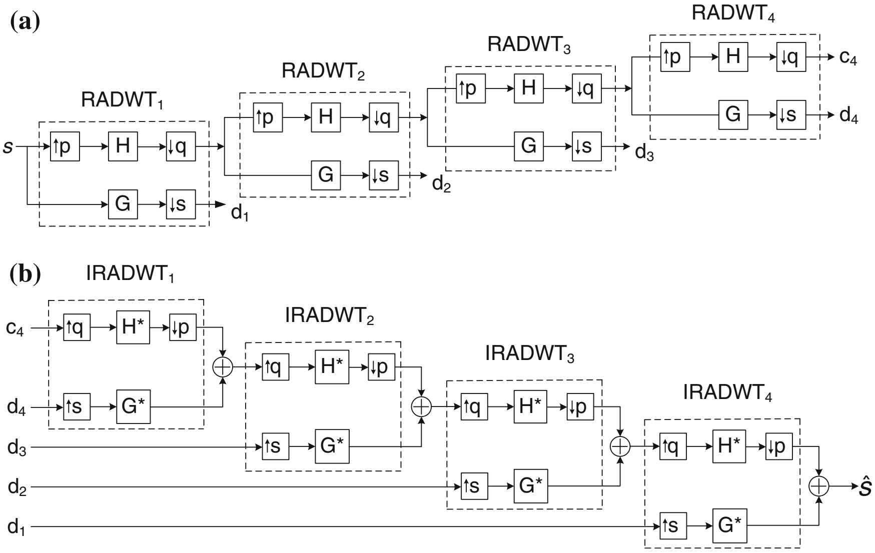


Fig. 49.3 Four stage RADWT decomposition. **a** Analysis filter bank. **b** Synthesis filter bank

where x and y are the coordinate vectors, $\mathrm{D_f}(\mathrm{x}, \mathrm{y})$ and $\mathrm{R_f}(\mathrm{x}, \mathrm{y})$ are domain and range filter components of the bilateral filter, respectively, which are defined as

$$\mathrm{D_f}(\mathrm{x,y}) = \exp\left[\frac{-\|\mathrm{y} - \mathrm{x}\|^2}{2\sigma_\mathrm{d}^2}\right] \text{and}\, \mathrm{R_f}(\mathrm{x}, \mathrm{y}) = \exp\left[\frac{-\|\mathrm{s(y)} - \mathrm{s(x)}\|^2}{2\sigma_\mathrm{r}^2}\right] \tag{49.6}$$

$\mathrm{N(x)}$ is the spatial neighborhood of s and C is the normalization constant defined as:

$$\mathrm{C} = \sum\nolimits_{\mathrm{y} \in \mathrm{N(x)}} \mathrm{D_f}(\mathrm{x}, \mathrm{y})\mathrm{R_f}(\mathrm{x}, \mathrm{y}) \tag{49.7}$$

where σ_d and σ_r are the domain and range parameters which control the behavior of the weights.

In bilateral filter, the choice of the parameters σ_d and σ_r are very important. If their values are very high, the filter behaves as a smoothing filter and will blur the edges. If their values are too low, the noise cannot be removed. The optimal value of σ_d is relatively insensitive to noise variance while the optimal value of σ_r varies significantly as the noise standard deviation [36]. The optimal value of range parameter σ_r depends on noise standard deviation linearly. In [36], an additive noise model based on bilateral filter has been already explained but if the filter is applied to speckle noise, the relationship between σ_r and noise variance will not be established because speckle noise is multiplicative in nature. In order to reduce the speckles in ultrasound images effectively, a modified bilateral filter has been presented by Tang et al. [37]. A speckle reducing filter with a modification in the expression of range filter has also been suggested by them and is given as:

$$R_f(x, y) = \exp\left[\frac{-\|s(y) - s(x)\|^2}{2\|s(x)\|^2 \sigma_r^2}\right] \tag{49.8}$$

In order to improve the effectiveness of the bilateral filter, an iterative bilateral filter is also presented by them.

49.2.3 *Thresholding (THR)*

Various thresholding schemes are provided in literature. These thresholding schemes provide the threshold coefficients by comparing the transformed coefficients against a threshold to remove the noise from a signal while preserving the important information of the original signal. The subband containing high frequency coefficients are processed with the thresholding techniques such as soft and hard thresholding but sometimes they carry the edges with large magnitude. The main task of thresholding approach is the selection of appropriate value of threshold (T). Now, thresholding is concentrated on NeighShrink thresholding (NST) which is improved by Zhou et al. [38]. The authors have performed various quantitative evaluations and proved that the NST performs better than other existing methods [38]. The performance of the RADWT-BLF is evaluated by denoising of the high frequency coefficients using soft and NeighShrink thresholding scheme.

49.2.3.1 Soft Thresholding

Soft thresholding is used to approximate the noisy detail coefficients $d(x, y)$ of the signal. The coefficients whose absolute values are lower than the particular threshold (T) are first set to zero and then scaling the nonzero coefficients i.e. whose values are greater than the threshold (T). So the noiseless coefficients are computed using soft thresholding as follows [23],

$$\mathrm{THR}(d(x, y)) = \begin{cases} d(x, y) - T; & d(x,y) > T \\ d(x, y) + T; & d(x, y) < -T \\ 0; & |d(x, y)| < T \end{cases} \tag{49.9}$$

T is defined as the universal threshold and it can be estimated as

$$T = \sigma_N \sqrt{2\log(N)} \tag{49.10}$$

where N is the length of the signal and σ_N is estimated standard deviation of noise which is computed using the median of absolute deviation from the RADWT coefficient (d_1). The σ_N is evaluated as

$$\sigma_N = \frac{\text{MEDIAN}(|d(x,y)|)}{0.6745} \tag{49.11}$$

49.2.3.2 NeighShrink Thresholding

To achieve the threshold coefficient, an improved NeighShrink thresholding (NST) algorithm is proposed that is based on the Stein's unbiased risk estimate (SURE) [38]. After getting the RADWT coefficients, a threshold value is required. For the RADWT coefficient $d(x,y)$ to be threshold, consider a square window $w(x,y)$ centered at the coefficients.

Let $S^2(x,y) = \sum_{(k,l)\in w(x,y)} d^2(k,l)$ and the thresholding expression is given as [38]

$$\widehat{d}(x,y) = \left(1 - \frac{T^2}{S^2(x,y)}\right)_+ d(x,y) \tag{49.12}$$

where $\widehat{d}(x,y)$ is the estimator of the unknown noiseless coefficients and thresholding factor $\emptyset = \left(1 - T^2/S^2(x,y)\right)_+$. Here the '+' sign means to keep only the positive values while it is set to zero when it is negative. The optimal value of threshold T and window size l is determined for every subband using SURE by minimizing the mean squared error or risk of the corresponding subband. For ease of notation, firstly arrange the noisy RADWT coefficients of different subband are arranged to a vector $d = \{d(x,y)\}$. Stein showed that for almost any fixed estimator $\widehat{d}$ based on the signal $d(x,y)$, the risk $E\left\{\left\|\widehat{d} - d\right\|^2\right\}$ can be estimated.

$$E\left\{\left\|\widehat{d} - d\right\|^2\right\} = E\{\text{SURE}(d,T,l)\} \tag{49.13}$$

where $\text{SURE}(d,T,l) = N + E\left\{\|G(d)\|^2 + 2\nabla \cdot G(d)\right\}$ and $G(d) = \{G(i)\}_{i=1}^{n} = \widehat{d} - d$, n is the number of the RADWT coefficients in a subband, arranged in 1D vector. This is the expected risk estimate on a particular subband for a square neighboring window. The optimal threshold T and neighboring window size l for different subband minimize $\text{SURE}(d,T,l)$. Accordingly,

$$(T,l) = \text{argminSURE}(d,T,l) \tag{49.14}$$

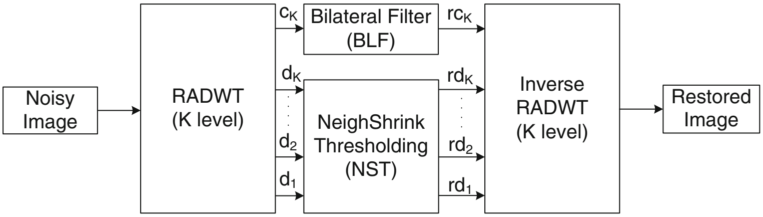


Fig. 49.4 Process flow of proposed method

49.3 Proposed Method

The proposed RADWT based denoising method consists of the three steps as shown in Fig. 49.4: (1) select the appropriate filter, number of decomposition level, dilation factors (p and q) and high pass sampling factor (s), to compute the RADWT noisy coefficients from the input noisy signal, (2) select and perform the appropriate filtering algorithm on these RADWT noisy coefficients, to get the modified RADWT coefficients, and (3) finally obtain the reconstructed signal by taking inverse RADWT of these manipulated RADWT coefficients.

Let $\mathrm{s(x,y)}$ be the input noisy image signal on which a rational-dilation wavelet transform is applied to obtain the noisy RADWT coefficients. After applying the RADWT at different levels, it is decomposed into a set of the RADWT coefficient as a vector consisting of K subbands with different spectral resolution such as

$$\mathrm{s^{RADWT}(x,y) = RADWT(s(x,y))} \tag{49.15}$$

where $\mathrm{s^{RADWT}(x,y)} = \xi = \{\delta_\mathrm{K}, \mathrm{c_K}\}$ and $\delta_\mathrm{K} = \{\mathrm{d_1, d_2, \ldots, d_K}\}$

High frequency noise mainly exists in the lower stage coefficient $(\mathrm{d_1})$. On the other hand, in the last stage coefficient $(\mathrm{c_K})$ noise is negligible. Thus, performing the appropriate filtering in each layer suppress the noise effectively without degrading the signal which are varying very slowly. The non-linear bilateral filter (BLF) is applied on the RADWT coefficients of the final stage i.e. low frequency component $\mathrm{c_K}$ and rest of all the RADWT coefficients i.e. δ_K is modified by the NST algorithm.

$$\mathrm{rc_K = BLF(c_K)\,and\,r\delta_K = NST(\delta_K)} \tag{49.16}$$

where $\mathrm{r\delta_K} = \{\mathrm{rd_{1,}rd_{2,}\ldots, rd_K}\}$, After applying these two filtering techniques, the modified filtered RADWT coefficients are obtained, which can be expressed as

$$\widehat{\mathrm{s}}^{\mathrm{RADWT}}\mathrm{(x,y)} = \mathrm{r}\xi = \{\mathrm{r}\delta_\mathrm{K}, \mathrm{rc_K}\} = \{\mathrm{rd_1, rd_2, \ldots, rd_K, rc_K}\} \tag{49.17}$$

The inverse RADWT is applied on the filtered coefficients to reconstruct the approximation of the noisy image signal i.e. $\widehat{\mathrm{s}}\mathrm{(x,y)}$.

$$\widehat{\mathrm{s}}\mathrm{(x,y)} = \mathrm{RADWT^{-1}}\left(\widehat{\mathrm{s}}^{\mathrm{RADWT}}\mathrm{(x,y)}\right) \tag{49.18}$$

49.4 Experimental Results and Discussions

To analyze the noise suppression and edge preservation capability, different performance indices have been used. Mean square error (MSE) and peak signal to noise ratio (PSNR) are the parameters to measure the noise suppression capability; however, they cannot be optimal with respect to perceived quality or reflect the performance of edge preservation capability of the denoising approaches. Hence the other assessment parameters such as image quality index (IQI) [39], correlation coefficients (RHO) [40], structural similarity index metrics (SSIM) [41], are used to measure the closeness of the fine details in denoised image with respect to the original image. For the quantitative evaluation of the edge preservation in denoised image, Pratt's figure of merit (FOM) [13] is most commonly used parameter with Canny operator.

To evaluate the performance of the proposed method, various experiments are being performed on different speckled ultrasound images. Test ultrasound images of bone fracture and other ultrasound images were acquired from the open source medical image database available at http://www.ultrasoundcases.info/Category.aspx?cat=73 and http://rad.usuhs.edu/medpix/parent.php3?mode=image_atlas. The analysis has been performed on several ultrasound images of bone fracture to evaluate the validity of the proposed method.

49.4.1 Experiment 1

To investigate the performance of the RADWT based proposed method, experiments are performed on thirty different ultrasound images of bone fracture in which few images are shown in Fig. 49.5a_1–c_1. The performance of the proposed method depends on the number of decomposition level (K), dilation factor (p and q) and high pass sampling factor (s) of RADWT decomposition, σ_d and σ_r of bilateral filter. The best optimal values of these parameters are determined by performing the successive experiments on the processed ultrasound images. The speckled ultrasound images are processed using proposed method with optimal values $K = 4$, $p = 3$, $q = 4$, $s = 2$ for RADWT decomposition, window size $= 11 \times 11$, $\sigma_d = 1.8$ and $\sigma_r = 2\sigma_n$ for performing bilateral filter.

The performance of the proposed method is shown in Fig. 49.5a_2–c_2 which shows the noise suppression as well as visualization capabilities of the proposed method.

Figures 49.6, 49.7 and 49.8 show the enhancement of the different ultrasound images of bone fracture using the proposed method and algorithm presented in [35]. Figure 49.6 is the rib fracture ultrasound image with cortical interruption of different samples. Figure 49.7 is ultrasound image of rib fracture with callus formation and Fig. 49.8 is fibula fracture. In all these figs., a–c show the different image samples which are taken to perform the experiments. '1' denotes the

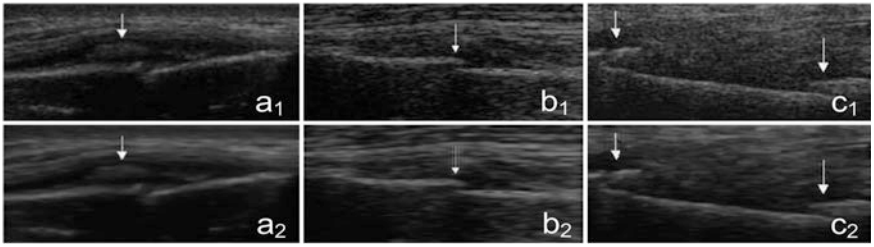


Fig. 49.5 *a_1–c_1* Test ultrasound images of bone fracture. *a_2–c_2* Denoised images processed with the proposed method

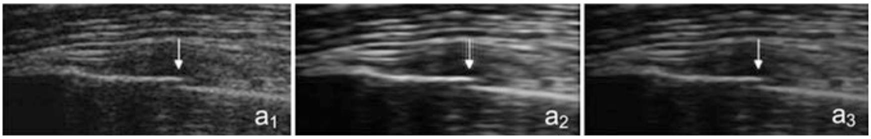


Fig. 49.6 Ultrasound image of rib fracture with cortical interruption

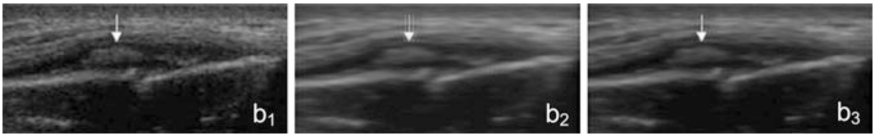


Fig. 49.7 Ultrasound image of rib fracture with callus formation

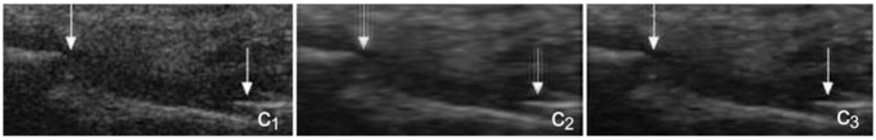


Fig. 49.8 Ultrasound image of fibula fracture with cortical interruption

speckled bone fracture ultrasound images, ‘2’ denote the algorithm presented in [35] and ‘3’ indicates the proposed method.

To evaluate the superiority of the proposed method, the qualitative and quantitative performance of the proposed method (M-2) is compared with the RADWT based thresholding method with bilateral filter (M-1) presented in [35]. Figures 49.6–49.8 show the comparative visual performance for bone fracture ultrasound images. Apart from the visual assessment, the quantitative evaluation of the proposed method has been also done. The comparative evaluation is done on the basis of different performance measures as mentioned above. This comparative quantitative performance is shown using the box plot.

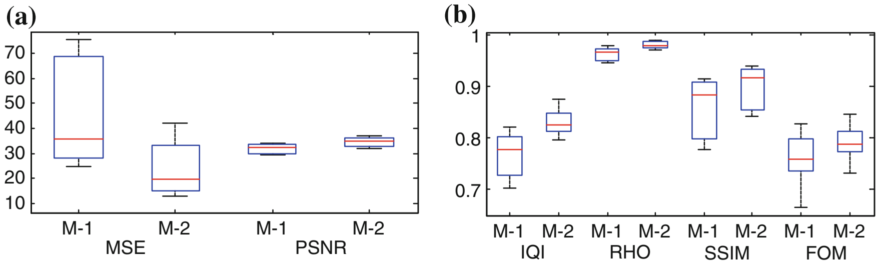


Fig. 49.9 Comparative performance for bone fracture images. **a** Noise suppression performance. **b** Quality, feature and edge preservation performance

Figure 49.9a shows the box plot of MSE and PSNR values for both methods. The top and bottom of each rectangular box indicates the 25th and 75th percentile, respectively, with the median shown inside the box. From the figure, it is clearly seen that the median values of the MSE and PSNR are lower and higher, respectively, for the method M-2. It indicates the superiority of the proposed method over the others one in terms of noise suppression.

Figure 49.9b indicates the box plot of the four different parameters IQI, RHO, SSIM and FOM values for both the methods. It is seen from Fig. 49.9b that the value of all the parameters is higher or closer to unity for the proposed method (M-2), which shows the better performance of the feature and edge preservation of the proposed method.

49.4.2 *Experiment 2*

To verify the despeckling capability of the proposed method, different ultrasound images are used rather than bone fracture images. There are fifty different ultrasound images which are processed with the different algorithms. Figure 49.10 shows the performance of the proposed method and compares with the algorithm presented in [35]. Figure 49.10, ‘1’, ‘2’ and ‘3’ denotes as per the above experiment.

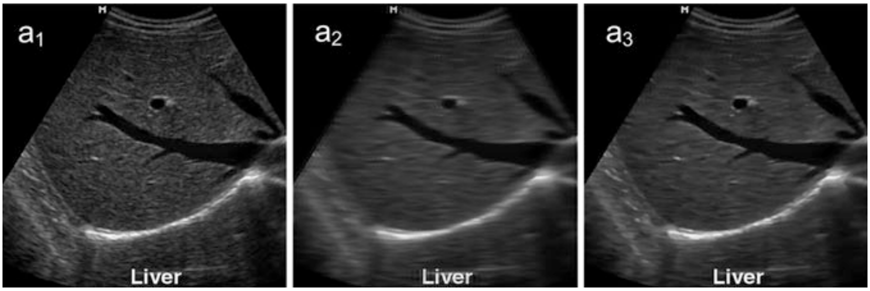


Fig. 49.10 Comparative performance for ultrasound image

Table 49.1 Performance comparison of denoising using different methods

Img.	Method	MSE	PSNR	IQI	RHO	SSIM	FOM
1	M-1	172.292	25.768	0.504	0.971	0.717	0.796
	M-2	97.022	28.262	0.624	0.984	0.84	0.802
2	M-1	75.876	29.33	0.567	0.972	0.792	0.656
	M-2	32.32	33.036	0.685	0.988	0.889	0.748
3	M-1	142.673	26.587	0.559	0.965	0.621	0.768
	M-2	89.708	28.603	0.564	0.978	0.714	0.79
4	M-1	65.354	29.978	0.575	0.978	0.813	0.792
	M-2	21.922	34.722	0.697	0.993	0.907	0.847
5	M-1	110.945	27.68	0.573	0.977	0.686	0.685
	M-2	61.187	30.264	0.657	0.987	0.756	0.742

The values of assessment parameters are listed in Table 49.1, for five different ultrasound images processed with both the methods M-1 and M-2. From the Table 49.1, it is clearly seen that the proposed method obviously outperforms the RADWT based thresholding method [35] in terms of all the assessment parameters as mentioned above. The proposed method provides the relatively stable PSNR gains over the others. This method has achieved approximately 11–12 % larger value of PSNR (in dB) as noise suppression capability, 15–16 % increased value in image quality and 6–7 % higher value of FOM as its edge preservation performance of the proposed method over the method presented in [35] for different ultrasound images.

49.5 Conclusion

In this chapter, RADWT based approach using non-linear bilateral filter has been presented to suppress the speckle noise from the real ultrasound images of bone fracture. In this proposed method, variations in frequency resolution features of rational-dilation wavelet transform are utilized and image is decomposed into different subbands at different stages. Bilateral filtering of the final stage coefficient suppresses the large amplitude noise components and NeighShrink thresholding provides the modified threshold coefficients which improves the denoising efficiency. The performance of the proposed method i.e. RADWT-BLF-NST is evaluated on large number of the ultrasound images of bone fracture and others also. It can be seen from the results that for almost all the denoised images used here with various noise levels, there is an improvement in MSE, PSNR, IQI, RHO, SSIM, and FOM as compared to the other method. Further, it can be observed also from the results that there is an improvement in terms of visual appearance of the despeckled regions of the ultrasound images. Further, the proposed algorithm not only shows improvement in the noise suppression in terms of quantitative measures such as MSE and PSNR, but also has the capability of edge preservation measured using the different parameter like FOM.

References

1. Dhawan AP (2003) Medical image analysis. Wiley Inc., New York
2. Lazović D, Wegner U, Peters G, Gossé F (1996) Ultrasound for diagnosis of apophyseal injuries. Knee Surg Sports Traumatol Arthrosc 3:234–237
3. Hübner U, Schlicht W, Outzen S, Barthel M, Halsband H (2000) Ultrasound in the diagnosis of fractures in children. J Bone Joint Surg Br 82(8):1170–1173
4. Rathfelder FJ, Paar O (1995) Possibilities for using sonography as a diagnostic procedure in fractures during the growth period. Der Unfallchirurg 98(12):645–649
5. Heiner JD, Proffitt AM, McArthur TJ (2011) The ability of emergency nurses to detect simulated long bone fractures with portable ultrasound. Int Emerg Nurs 19(3):120–124
6. Heiner JD, McArthur TJ (2009) A simulation model for the ultrasound diagnosis of long-bone fractures. Simul Healthc 4(4):228–231
7. Marshburn TH, Legome E, Sargsyan A, Li SMJ, Noble VA, Dulchavsky AS, Sims C, Robinson D (2004) Goal-directed ultrasound in the detection of long-bone fractures. J Trauma Acute Care Surg 57(2):329–332
8. Elamvazuthi I, Zain MLBM, Begam KM (2013) Despeckling of ultrasound images of bone fracture using multiple filtering algorithms. Math Comput Model 57(1–2):152–168
9. Bitschnau R, Gehmacher O, Kopf A, Scheier M, Mathis G (1996) Ultrasonography in the diagnosis of rib and sternal fracture. Eur J Ultrasound 3(2):197–297
10. Griffith JF, Rainer TH, Ching AS, Law KL, Cocks RA, Metreweli C (1999) Sonography compared with radiography in revealing acute rib fracture. Am J Roentgenol 173(6): 1603–1609
11. Hurley ME, Keye GD, Hamilton S (2004) Is ultrasound really helpful in the detection of rib fractures? Injury 35(6):562–566
12. Mittal D, Kumar V, Saxena SC, Khandelwal N, Karla N (2010) Enhancement of the ultrasound images by modified anisotropic diffusion method. Biol Eng Comput 48(12): 1281–1291
13. Pratt WK (2006) Digital image processing. Wiley, New York
14. Loupas T (1989) An adaptive weighted median filter for speckle suppression in medical ultrasonic images. IEEE Trans Circuits Syst 36(1):129–135
15. Gonzalez RC, Woods RE (2001) Digital image processing. Prentice-Hall, Englewood Cliffs
16. Perona P, Malik J (1990) Scale space and edge detection using anisotropic diffusion. IEEE Trans Pattern Anal Mach Intell 12(7):629–639
17. Kuan DT, Sawchuk AA, Strand TC, Chavel P (1985) Adaptive noise smoothing filter for images with signal dependent noise. IEEE Trans Pattern Anal Mach Intell 7(2):165–177
18. Lee JS (1980) Digital image enhancement and noise filtering by use of local statistics. IEEE Trans Pattern Anal Mach Intell 2(2):165–168
19. Yu Y, Acton ST (2002) Speckle reducing anisotropic diffusion. IEEE Trans Image Process 11(11):1260–1270
20. Liu X, Liu J, Xu X, Chun L, Deng Y (2011) A robust detail preserving anisotropic diffusion for speckle reduction in ultrasound images. BMC Genomics 12:1–10
21. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biometrika 81(3):425–455
22. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90:1200–1224
23. Donoho DL (1995) De-noising by soft thresholding. IEEE Trans Inf Theory 41(3):613–627
24. Gupta S, Chauhan RC, Saxena SC (2004) Wavelet based statistical approach for speckle reduction in medical ultrasound images. Med Biol Eng Comput 42:189–192
25. Foder IK, Kamath C, Kamath R (2001) Denoising through wavelet shrinkage: an empirical study. J Electron Imag 12:151–160
26. Achim A, Bezerianos A, Tsakalides P (2001) Novel bayesian multiscale method for speckle removal in medical ultrasound images. IEEE Trans Med Imag 20(8):772–783

27. Chang SG, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process 9(9):1532–1546
28. Michailovich OV, Tannenbaum A (2006) Despeckling of medical ultrasound images. IEEE Trans Ultrason Ferroelectr Freq Control 53(1):64–78
29. Bhutada GG, Anand RS, Saxena SC (2010) Fast adaptive learning algorithm for sub-band adaptive thresholding function in image denoising. Int J Comput Intell Stud 1(3):227–241
30. Abrahim BA, Kadah Y (2011) Speckle noise reduction method combining total variation and wavelet shrinkage for clinical ultrasound imaging. In: Proceedings of 1st middle east conference biomedical engineering (MECBME), pp 80–83
31. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Proceedings of sixth international conference computer vision, pp 839–846
32. Vanithamani R, Umamaheswari G (2011) Wavelet based despeckling of medical ultrasound images with bilateral filter. In: Proceedings of IEEE region 10 Conference TENCON, pp 389–393
33. Anand CS, Sahambi JS (2010) Wavelet domain non-linear filtering for MRI denoising. Magn Reson Imaging 28(6):842–861
34. Bayram I, Selesnick IW (2009) Frequency-domain design of overcomplete rational-dilation wavelet transforms. IEEE Trans Signal Process 57(8):2957–2972
35. Gupta D, Anand RS, Tyagi B (2012) Enhancement of medical ultrasound images using non-linear filtering based on rational-dilation wavelet transform, lecture notes in engineering and computer science. In: Proceedings of the world congress on engineering and computer science 2012, WCECS 2012, San Francisco, USA, 24–26 Oct 2012, pp 615–620
36. Zhang M, Gunturk BK (2008) Multiresolution bilateral filtering for image denoising. IEEE Trans Image Process 17(12):2324–2332
37. Tang J, Guo S, Sun Q, Deng Y, Zhou D (2010) Speckle reducing bilateral filter for cattle follicle segmentation. BMC Genomics 11(2):1–9
38. Denweng Z, Wengang C (2008) Image denoising with an optimal threshold and neighbouring window. Pattern Recognit Lett 29:1694–1697
39. Wang Z, Bovik AC (2002) A universal image quality index. IEEE Signal Process Lett 9(3):81–84
40. Thakur A, Anand RS (2005) Image quality based comparative evaluation of wavelet filters in ultrasound speckle reduction. Digit Signal Process 15:455–465
41. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612

Index

H. K. Kim et al. (eds.), *IAENG Transactions on Engineering Technologies*,
Lecture Notes in Electrical Engineering 247, DOI: 10.1007/978-94-007-6818-5,