LECTURE NOTES IN ELECTRICAL ENGINEERING 11

Nikos Mastorakis
John Sakellaris
Editors

Advances in Numerical Methods

Springer

Advances in Numerical Methods

Lecture Notes in Electrical Engineering

Volume 11

Advances in Numerical Methods
Nikos Mastorakis and John Sakellaris
978-0-387-76482-5

Radio Monitoring: Problems, Methods, and Equipment
Anatoly Rembovsky, Alexander Ashikhmin, Vladimir Kozmin, and Sergey Smolsky
978-0-387-98099-7

Incorporating Knowledge Sources into Statistical Speech Recognition
Sakti, Sakriani, Markov, Konstantin, Nakamura, Satoshi, and Minker, Wolfgang
978-0-387-85829-6

Intelligent Technical Systems
Martínez Madrid, Natividad; Seepold, Ralf E.D. (Eds.)
978-1-4020-9822-2

Languages for Embedded Systems and their Applications
Radetzki, Martin (Ed.)
978-1-4020-9713-3

Multisensor Fusion and Integration for Intelligent Systems
Lee, Sukhan; Ko, Hanseok; Hahn, Hernsoo (Eds.)
978-3-540-89858-0

Designing Reliable and Efficient Networks on Chips
Murali, Srinivasan
978-1-4020-9756-0

Trends in Communication Technologies and Engineering Science
Ao, Sio-Iong; Huang, Xu; Wai, Ping-kong Alexander (Eds.)
978-1-4020-9492-7

Functional Design Errors in Digital Circuits: Diagnosis Correction and Repair
Chang, Kai-hui, Markov, Igor, Bertacco, Valeria
978-1-4020-9364-7

Traffic and QoS Management in Wireless Multimedia Networks: COST 290 Final Report
Koucheryavy, Y., Giambene, G., Staehle, D., Barcelo-Arroyo, F., Braun, T., Siris,V. (Eds.)
978-0-387-85572-1

Proceedings of the 3rd European Conference on Computer Network Defense
Siris, V.; Ioannidis, S.; Anagnostakis, K.; Trimintzios, P. (Eds.)
978-0-387-85554-7

Continued after index

Nikos Mastorakis • John Sakellaris
Editors

Advances in Numerical Methods

Springer

Editors

Nikos Mastorakis
Hellenic Naval Academy
Military Institutes of University Education
Piraeus 185 39
Greece
mastor@hna.gr

John Sakellaris
Technological Educational Institute
of Thessaloniki
Thessaloniki 601 00
Greece
jsakel@central.ntua.gr

ISBN 978-0-387-76482-5 e-ISBN 978-0-387-76483-2
DOI 10.1007/978-0-387-76483-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009928455

Printed on acid-free paper

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

Table of Contents

Part 1

Applied mathematics

Part 1 of this book includes a selection of papers from various conferences sponsored by WSEAS. The society WSEAS organizes and sponsors more than 70 conferences per year and publishes several scientific journals. The chapters of this volume cover a wide area of applied mathematics and their applications in science and technology and have been written by prominent invited lecturers. This part of the book attempts to give a panoramic view of the most promising areas of modern applied mathematics (functional analysis, partial differential equations, ordinary differential equations, numerical analysis, probabilities and statistics, control, etc.) in problems of mathematical physics and engineering. The chapters of this volume are actually extended versions of papers presented in many WSEAS conferences, but have additionally passed by a second round of strict review after the conferences. We thank the reviewers for their difficult task as well as we thank the authors for their patience until the final edition of this book. Finally, we wish to thank Springer Verlag for the excellent collaboration during the editing and publishing process.

Chapter 1

Similarity solutions of an MHD boundary-layer flow of a non-Newtonian fluid past a continuous moving surface

M. Guedda,[1] Z. Mahani,[2] M. Benlahcen,[1] A. Hakim[2]

[1]Faculty of Mathematics, University Picardie Jules-Vernes, 33 Rue St Leu 80 039 Amiens, France, Mohamed.guedda@u-picardie.fr
[2]Department of Mathematics and Informatics, University Cadi Ayad, P.O Box 549 Marrakech, Morocco, zouhir_mahani@hotmail.com

Abstract. This chapter deals with a theoretical and numerical analysis of similarity solutions of the 2-D boundary-layer flow of a power-law non-Newtonian fluid past a permeable surface in the presence of a magnetic field $B(x)$ applied perpendicular to the surface. The magnetic field B is assumed to be proportional to $x^{(m-1)/2}$, where x is the coordinate along the plate measured from the leading edge and m is a constant. The problem depends on the power-law exponent m, the power-law index n and the magnetic parameter M or the Stewart number. It is shown, under certain circumstance, that the problem has an infinite number of solutions.

Keywords. Similarity solutions, Boundary-layer flow, Non-Newtonian fluid

1.1 Introduction

The prototype of the problem under investigation is

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_1,

$$\alpha\frac{\partial}{\partial y}\left(\left|\frac{\partial^2\psi}{\partial y^2}\right|^{n-1}\frac{\partial^2\psi}{\partial y^2}\right)+\frac{\partial\psi}{\partial x}\frac{\partial^2\psi}{\partial y^2}-\frac{\partial\psi}{\partial y}\frac{\partial^2\psi}{\partial xy}+u_e\frac{\partial u_e}{\partial x}-\sigma B^2\left(\frac{\partial\psi}{\partial y}-u_e\right)=0 \tag{1.1}$$

with the boundary conditions

$$\frac{\partial\psi}{\partial y}(x,0)=u_w(x),\quad \frac{\partial\psi}{\partial x}(x,0)=v_w(x) \tag{1.2}$$

and

$$\lim_{y\to\infty}\frac{\partial\psi}{\partial y}(x,0)=u_e(x) \tag{1.3}$$

where the unknown function is the stream function ψ, u_e is the free stream velocity, α, k, σ and n are kinematic viscosity, permeability, electric conductivity and power-law index, respectively. The above problem is a model for the first approximation to 2-D laminar incompressible flow of an electrically conducting non-Newtonian power-law fluid past a moving plate surface. Here the $x \geq 0$ and $y \geq 0$ are the Cartesian coordinates along and normal to the plate with $y = 0$ being the plate and the plate origin located at $x = y = 0$. The magnetic field is given by $B(x) = B0x^{(m-1)/2}$, $B0 > 0$, and is assumed to be applied normally to the surface.

The problem in Eqs. (1.1), (1.2) and (1.3) is deduced from the boundary-layer approximation

$$\frac{\partial u}{\partial x}+\frac{\partial v}{\partial y}=0 \tag{1.4}$$

and

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}=\alpha\frac{\partial}{\partial y}\left(\left|\frac{\partial u}{\partial y}\right|^{n-1}\frac{\partial u}{\partial y}\right)+u_e\frac{\partial u_e}{\partial x}-\sigma B^2(u-u_e) \tag{1.5}$$

and

$$u(x,0)=u_w(x),\quad v(x,0)=v_w(x) \qquad (1.6)$$
$$\lim_{y\to\infty}u(x,y)=u_e(x)$$

According to $u=\left(\frac{\partial\psi}{\partial y}\right)$ and $v=-\left(\frac{\partial\psi}{\partial x}\right)$, where u and v represent the components of the fluid velocity in the direction of increasing x and y.

Here, it is assumed that the flow behaviour of the non-Newtonian fluid is described by the Ostwald–de Waele power-law model, where the shear stress is related to the strain rate $\left(\frac{\partial u}{\partial y}\right)$ by the expression [7, 13, 20]

$$\tau = K\left|\frac{\partial u}{\partial y}\right|^{n-1}\frac{\partial u}{\partial y}$$

where K is a positive constant and $n > 0$ is called the power-law index. The case $n < 1$ is referred to as pseudo-plastic fluids (or shear-thinning fluids), the case $n > 1$ is known as dilatant or shear-thickening fluids. The Newtonian fluid is, of course, a special case where the power-law index n is one. The stretching, suction/injection velocities and the free stream velocity are assumed to be of the form

$$u_w(x) = u_w x^m, \quad v_w(x) = -v_s x^{\frac{m(2n-1)-n}{n+1}} \qquad (1.7)$$
$$u_e(x) = u_\infty x^m$$

where u_w and u_∞ are positive constants and v_s is a real number with $v_s < 0$ for injection and $v_s > 0$ for suction.

The magnetohydrodynamic (MHD) flow problems find applications in many physical, geophysical and industrial fields. Pavlov [17] was the first who examined the MHD flow over a stretching wall in an electrically conducting fluid, with a uniform magnetic field. Further studies in this direction are those of Chakrabarti and Gupta [8], Vajravelu [26], Takhar et al. [25, 22], Kumari et al. [14], Andersson et al. [3] and Watanabe and Pop [27]. The possibility of obtaining similarity solutions for the MHD flow over a stretching permeable surface subject to suction or injection was considered by [8, 26] for some values of the mass transfer parameter, say, f_w, and by Pop and Na [18], for large values of f_w and where the stretching velocity varies linearly with the distance and where the suction/injection velocity is constant. The MHD flow over a stretching permeable surface with variable suction/injection velocity can be found in [9]. A complete physical interpretation of the problem can be found in [8, 19, 21, 24].

In this chapter, we will examine similarity solutions to Eqs. (1.1), (1.2) and (1.3) in the usual form

$$\psi(x,y) = \lambda x^s f(\eta), \quad \eta = \gamma \frac{y}{x^r} \qquad (1.8)$$

where s and r are real numbers, $\lambda>0$ and $\gamma>0$ are such that

$$\lambda\gamma = u_\infty, \quad \alpha\lambda^{n-2}\gamma^{2(n-1)} = 1$$

Using Eqs. (1.1) and (1.8) we find that the profile function satisfies

$$\left(\left|f''\right|^{n-1} f''\right)' + sff'' + m\left(1 - f'^2\right) + M\left(1 - f'\right) = 0 \tag{1.9}$$

if and only if
m=$s-r$, $s(2-n)$+$r(2n-1)$
which leads to

$$s = \frac{1 + m(2n-1)}{1+n}$$

In Eq. (1.9) the primes denote differentiations with respect to the similarity variable $\eta \in (0,\infty)$, the unknown function f denotes the similar stream function and its derivative, after suitable normalisation, represents the velocity parallel to the surface. The parameter $M = \left(\dfrac{\sigma B_0^2}{u_\infty \rho}\right)$ is the magnetic parameter. Equation (1.9) will be solved subject to the boundary conditions

$$f(0) = a, \quad f'(0) = b \tag{1.10}$$

and

$$f'(\infty) = \lim_{\eta\to\infty} f'(\eta) = 1 \tag{1.11}$$

The parameters a and b are given by $a = (n+1)v_s\left(\alpha u_\infty^{2n-1}\right)^{\frac{-1}{n+1}}$ and $b = \left(\dfrac{u_w}{u_\infty}\right)$. For the Newtonian fluid (n=1), the ODE reads

$$f''' + sff'' + m\left(1 - f'^2\right) + M\left(1 - f'\right) = 0 \tag{1.12}$$

$$s = \frac{m+1}{2}$$

Numerical and analytical solutions to Eq. (1.12), in the absence of the free stream function ($f'(\infty) = 0$), were obtained in [13, 11, 18, 23]. Numerical solutions, in the presence of the free stream velocity, can be found in [4, 19, 24], for both momentum and heat transfers.

In a physically different but mathematically identical context, Eq. (1.12), with M = −m, which reads (by a scaling)

$$f''' + (1+m)ff'' + 2m(1-f') = 0 \tag{1.13}$$

has been investigated by Aly et al. [2], Brighi et al. [5], Brighi and Hoernel [6], Guedda [12], Magyari and Aly [15] and Nazar et al. [16]. This equation with the boundary condition ($a = 0$, $b = 1 +\varepsilon$)

$$f(0) = 0, \quad f'(0) = 1+\varepsilon, \quad f'(\infty) = 1 \tag{1.14}$$

arises in modelling the mixed convection boundary-layer flow in a porous medium. In [2] it is found that if m is positive and f″ takes place in the range $[\varepsilon_0,\infty)$, for some negative ε_0, there are two numerical solutions. The case $-1 \leq m \leq 0$ is also considered in [2]. The authors studied the problem for $\varepsilon_c \leq \varepsilon \leq 0.5$, for some $\varepsilon_c < 0$. It is shown that there exists ε_t such that the problem has two numerical solutions for $\varepsilon_c \leq \varepsilon \leq \varepsilon_t$. In [12] Guedda has investigated the theoretical analysis of Eqs. (1,13) and (1.14). It was shown that, if $-1 < m < 0$ and $-1 < \varepsilon < 1/2$, there is an infinite number of solutions, which indeed motivated the present work. Some new interesting results on the uniqueness of concave and convex solutions to Eqs. (1.13) and (1.14), for $m > 0$ and $\varepsilon > -1$, were reported in [6].

Most recently Aly et al. [1] have investigated the numerical and theoretical analysis of the existence, the uniqueness and non-uniqueness of solutions to Eqs. (1.13) and (1.14). It is shown that the problem has a unique concave solution and a unique convex solution for any $m > 0$ and $M \geq 0$. The case where the free stream is being retarded (increasing pressure) is also considered. The authors proved that, for any $-1/3 < m < -M < 0$ and any real number a, the problem (Newtonian case) has an infinite number of solutions. The multiplicity of solutions is also examined for $-1/2 < m < -M < 0$ provided $b > M/(m+1)$ and $a \geq b\sqrt{(m+1)b-M}$

The purpose of this note is to examine the problem of Eqs. (1.9), (1.10) and (1.11) for $-1 < m < -M < 0$.

1.2 Existence of infinitely many solutions

The interest in this section will be in the existence question of multiple solutions of problems (1.9),(1.10) and (1.11), where $-1 < m(2n-1)$, $m < 0$, and $m + M < 0$. The existence result will be established by means of a shooting method. Hence, the boundary condition at infinity is replaced by the condition

$$f''(0)=\tau \tag{1.15}$$

where τ is the shooting parameter which has to be determined. Local in η solution to Eqs. (1.9), (1.10) and (1.15) exists for every $\tau \in \Re$, and it is unique. Denote this solution by f_τ. Let us describe what conditions will be imposed for f_τ to be global and satisfies Eq. (1.11). Note that the real number τ has a physical meaning. This parameter originates from the local skin friction coefficient, c_f, and the local Reynolds numbers Re_x

$$\frac{1}{2}c_f Re_x^{1/(n+1)}=\left[\frac{m(2n-1)+1}{n(n+1)}\right]^{n/(n+1)}$$

where $Re_x=((u_w(x)^{2-n}x^n)/\alpha K)$.

Returning to the initial value problem of Eqs. (1.9), (1.10) and (1.15), our purpose is to derive favourable conditions on m, a and b such that f_η is global and satisfies $f_\tau'(\infty)=1$. We shall impose the condition $m\in(-1,0)$. The local solution f_τ satisfies the following equality that will be useful later on:

$$\left|f_\tau''(\eta)\right|^{n-1} f_\tau''(\eta)+sf_\tau'(\eta)f_\tau(\eta)-Mf_\tau(\eta)= \tag{1.16}$$

$$|\tau|^{n+1}\tau+sab-Ma-(M+m)\tau+\frac{1+3nm}{n+1}\int_0^\eta f_\tau'(s)^2\,ds$$

for all $0\le\eta<\eta_\tau$, where $(0,\eta_\tau)$ is the maximal interval of existence. Let us note that if η_τ is finite the function f_τ is unbounded on $(0,\eta_\tau)$ [1, 10].

Define

$$\Gamma=-\frac{3M}{4m}\left[1+\sqrt{1+\frac{16}{3}\frac{m}{M^2}(M+m)}\right]>1$$

where $M>0$ and $m+M<0$. Our main result is the following:

Theorem 1 Let $M>0$, $-1<m\ (2n-1)$ and $m<-M$. Assume $a\ge 0$ and $b\in(0,\Gamma)$. For any $\tau\in\Re$ such that

$$\tau^{n+1} \le (n+1)\left[\frac{1}{3}mb^3 + \frac{1}{2}Mb^2 - (M+m)b\right] \tag{1.17}$$

f_τ is global and satisfies

$$f'(\infty) = \lim_{\eta\to\infty} f'(\eta) = 1$$

Note that, since τ is arbitrary, problem of Eqs. (1.9), (1.10) and (1.11) has an infinite number of solutions. To prove Theorem 1 we use an idea given in [12]. First we have the following result.

Lemma 1 For any $a \ge 0$, $0 < \mathrm{b} < \Gamma$ and satisfying condition of Eq. (1.17), the function f_τ is positive, monotonic increasing on $(0, \eta_\tau)$ and global.

Moreover $f_\tau(\eta)$ tends to infinity with η and $\lim_{\eta\to\infty} f_\tau''(\eta) = 0$.

Proof

From Eq. (1.9) one sees

$$E' = -sf_\tau f_\tau''^2$$

on $(0,\eta_\tau)$, where E is the "Lyapunov" function for f_τ defined by

$$E = \frac{1}{n+1}\left|f_\tau''\right|^{n+1} - \frac{m}{3}f_\tau'^3 - \frac{M}{2}f_\tau'^2 + (M+m)f_\tau'$$

On the other hand, since $a \ge 0$ and $b > 0$ we may assume $f_\tau, f_\tau' > 0$ on some $(0,\eta_0)$, $0 < \eta_0 < \eta_\tau$.

Hence, the function E is monotonic decreasing on $(0,\eta_0)$. This implies

$$E(\eta_0) \le E(0) \tag{1.18}$$

which shows that $E(\eta_0) \le 0$, thanks to Eq. (1.17). If $f_\tau'(\eta_0)=0$, we get $\mathrm{E}(\eta_0)=\mathrm{E}(0) = 0$, and then $E(\eta) = 0$ for all $0 < \eta < \eta_0$. Therefore, $f_\tau'' = 0$ on $(0,\eta_0)$, and this implies $\eta = 0$ and $\mathrm{b} = 0$ or $b = \Gamma$, a contradiction. Hence f_τ is monotonic strictly increasing.

To show that f_τ is global, we use again the function E to deduce

$$\frac{1}{n+1}\left|f_\tau''\right|^{n+1}-\frac{m}{3}f_\tau'^{3}-\frac{M}{2}f_\tau'^{2}+(M+m)f_\tau'\leq \qquad (1.19)$$
$$\frac{1}{n+1}|\tau|^{n+1}-\frac{m}{3}b^3-\frac{M}{2}b^2+(M+m)b$$

Therefore, f_τ'' and f_τ' are bounded. Hence, f_τ is bounded on (0, η_τ), if η_τ is finite, which is absurd. Consequently $\eta_\tau = \infty$, that is, f_τ is global. Moreover, f_τ has a limit, say $L \in (0,\infty]$, at infinity, since f_τ' is positive. To demonstrate that L is infinite, we assume for the sake of contradiction that $L < \infty$. Hence, there exists a sequence (η_r) converging to infinity with r such that $f_\tau'(\eta_r)$ tends to 0 as n tends to infinity. Clearly,

$$-\frac{m}{3}f_\tau'(\eta_r)^3-\frac{M}{2}f_\tau'(\eta_r)^2+(M+m)f_\tau'(\eta_r)\leq E(\eta_r)\leq E(0),\quad \forall n\in \aleph$$

which implies $0 \leq E(\infty) \leq E(0)$. As above, we get a contradiction. It remains to show that the second derivative of f_τ tends to 0 at infinity, which is the case if f_τ'' is monotone on some interval $[\eta_0,\infty)$, since f_τ'' and f_τ' are bounded. Assume that $|f_\tau''|^{n-1}f_\tau''$ is not monotone on any interval $[\eta_0, \infty)$. Then, there exists an increasing sequence (η_r) going to infinity with r, such that $(|f_\tau''|^{n-1}f_\tau'')'(\eta_r) = 0$, $|f_\tau''|^{n-1}f_\tau''(\eta_{2r})$ is a local maximum and $|f_\tau''|^{n-1}f_\tau''(\eta_{2r}+1)$ is a local minimum. Setting $\eta = \eta_r$ in Eq. (1.9) yields

$$sf_\tau''(\eta_r)=-\frac{m\left(1-f_\tau'(\eta_r)^2\right)+M\left(1-f_t'(\eta_r)\right)}{f_\tau(\eta_r)} \qquad (1.20)$$

Because f_τ' is bounded and $f(\eta)$ tends to infinity with η, we get from Eq. (1.20) $f_\tau''(\eta_r) \to 0$ as $n \to \infty$, and (then) $f_\tau''(\eta) \to 0$ as $\eta \to \infty$.

In the next result we shall prove that $f_\tau'(\eta)$ goes to 1 as η approaches infinity and this shows that problem of Eqs. (1.9), (1.10) and (1.11) has an infinite number of solutions.

Lemma 2 Let f_τ be the (global) solution of Eqs. (1.9), (1.10) and (1.15) obtained in Lemma 1. Then

$$\lim_{\eta\to\infty}f_\tau'(\eta)=1$$

Proof

First we show that f_τ' has a finite limit at infinity. From the proof of Lemma 1 the function E has a finite limit at infinity, E_∞, say, and this limit takes place in the interval [$(4m+3M)/6$, 0]. Since f_τ'' goes to 0, we deduce that $-(m/3)f_\tau'^3-(m/2)f_\tau'^2+(M+m)f_\tau'$ tends to E_∞ as η → ∞. Let L_1 and L_2 be two nonnegative numbers given by

$$L_1=\liminf_{\eta\to\infty} f_\tau'(\eta) \qquad \text{and} \qquad L_2=\limsup_{\eta\to\infty} f_\tau'(\eta)$$

and satisfy

$$E_\infty=-\frac{m}{3}L_i^3-\frac{M}{2}L_i^2+(M+m)L_i \qquad i=1,2$$

Suppose that $L_1 \neq L_2$ and fix L so that $L_1 < L < L_2$. Let $(\eta_r)_n$ be a sequence tending to infinity with n such that $\lim_{\eta\to\infty} f_\tau'(\eta_r)=L$. Using the function E we infer

$$E_\infty=-\frac{m}{3}L^3-\frac{M}{2}L^2+(M+m)L$$

For all $L_1 < L < L_2$, which is impossible. Then $L_1 = L_2$. Hence, $f_\tau'(\eta)$ has a finite limit at infinity. Let us note this limit by L, which is nonnegative. Assume that $L = 0$. Then $E_\infty = 0$. Since E is a decreasing function, we deduce $E=0$
and get a contradiction. Hence $L > 0$. Next, we use identity of Eq. (1.16) to deduce, as η approaches infinity,

$$\left|f_\tau''\right|^{n-1} f_\tau''(\eta)=-(M+m)\eta+ML\eta-sL^2\eta+\frac{1+3nm}{n+1}L^2\eta+o(1)$$

$$\left|f_\tau''\right|^{n-1} f_\tau''(\eta)=\left[mL^2+ML-(M+m)\right]\eta+o(1)$$

And this is only satisfied if $mL^2 + ML - (M + m) = 0$, which implies $L = 1$, since L is positive.

This ends the proof of the lemma and the proof of Theorem 1.

Lemma 2 shows also that $E_\infty=(4m+3M)/6<0$. We finish this chapter by a nonexistence result in the case $m(2m-1)\le -1$, $n > 1/2$ and $b \ge \Gamma$.

Theorem 2 The problem of Eqs. (1.9) and (1.10) has no nonnegative solution for $M > 0$, $m < -M$, $m(2n-1) < -1$ and $b \geq \Gamma$.

Proof

Let f be a nonnegative solution to eqs. (1.9) and (1.10). As above, the function E satisfies

$$E' = -\frac{1+m(2n-1)}{n+1} ff''^2,$$

which is nonnegative.

Clearly, $E(0) \leq \lim_{t\to\infty} E(t)$ hence

$$-\frac{m}{3}b^3 - \frac{M}{2}b^2 + (M+m)b \leq \frac{4m+3M}{6} < 0$$

and this is not possible.

1.3 Numerical results

Now we present the numerical results for different values of n, m and M:

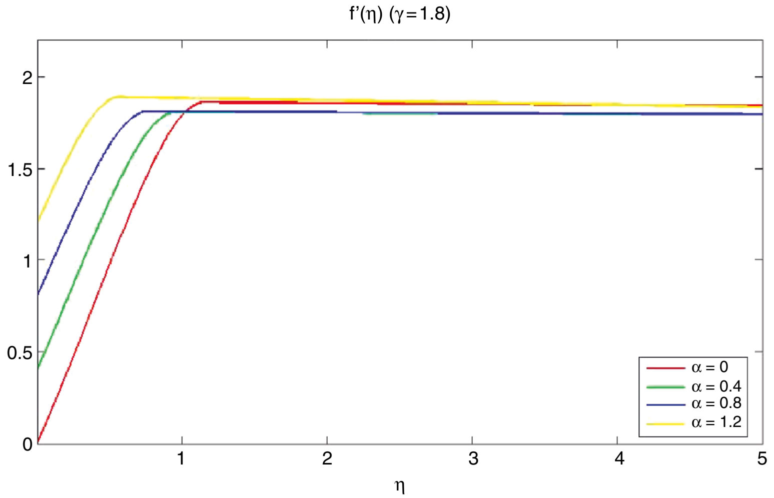


Fig. 1.1. $f'(\eta)$ for M=1.2, m=−1.5, γ=1.8 and some values of α

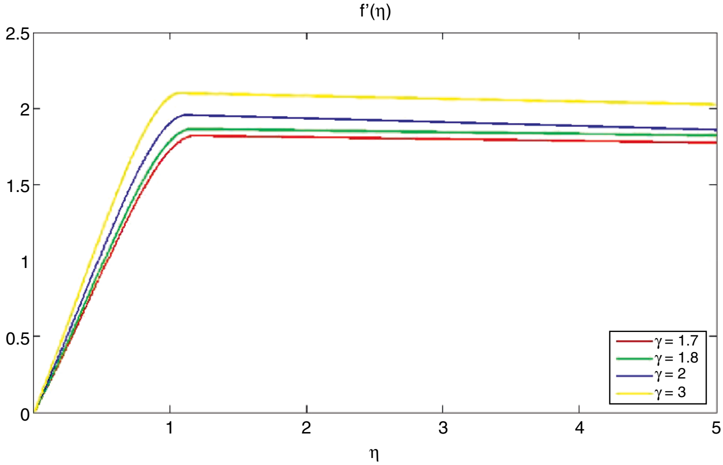


Fig. 1.2. $f'(\eta)$ for M=1.2, m=− 1.5, α=0 and some values of γ

Acknowledgments. The authors wish to thank Robert Kersner for interesting discussion. This work was partially supported by Direction des Affaires Internationals (UPJV) Amiens, France, and by PAI No MA/05/116.

References

1. Aly EH, Amkadni M, Guedda M (In preparation) A note on MHD flow over a stretching permeable surface
2. Aly EH, Elliott L, Ingham DB (2003) Mixed convection boundary-layer flow over a vertical surface embedded in a porous medium. Eur J Mech B Fluids 22:529–543
3. Andersson HI (1992) MHD flow of a viscous fluid past a stretching surface. Acta Mechanica 95:227–230
4. Anjali Devi SP, Kandasamy R (2003) Thermal stratification effects on non linear MHD laminar boundary-layer flow over a wedge with suction or injection. Int Comm Heat Mass Transf 30:717–725
5. Brighi B, Benlahsen M, Guedda M, Peponas S (In preparation) Mixed convection on a wedge embedded in a porous medium
6. Brighi B, Hoernel D (2006) On the concave and convex solution of mixed convection boundary layer approximation in a porous medium. Appl Math Lett 19:69–74
7. Callegari AJ, Frieddman MB (1968) An analytical solution of a nonlinear, singular boundary value problem in the theory of viscous fluids. J Math Analy Appl 21:510–529
8. Chakrabarti A, Gupta AS (1979) Hydromagnetic flow and heat transfer over a stretching sheet. Quart Appl Math 37:73–78

9. Chaturvedi N (1996) On MHD flow past an infinite porous plate with variable suction. Ener Conv Mgmt 37(5):623–627
10. Coppel WA (1960) On a differential equation of boundary layer theory. Phil Trans Roy Soc London Ser A 253:101–136
11. Cortel R (2005) Flow and heat transfer of an electrically conducting fluid of second grade over a stretching sheet subject to suction and to a transverse magnetic field. Int J Heat Mass Transf (published online)
12. Guedda M (2006) Multiple solutions of mixed convection boundary-layer approximations in a porous medium. Appl Math Lett 19:63–68
13. Howell TG, JENG DR, De Witt KJ (1997) Momentum and heat transfer on a continuous moving surface in a power-law fluid. Int J Heat Mass Transfer 40(8):1853–1861
14. Kumari M, Takhar HS, Nath G (1990) MHD flow and heat transfer over a stretching surface with prescribed wall temperature or heat flux. Warm und Stoffubert 25:331–336
15. Magyari E, Aly EH (2006) Exact analytical solution for a thermal boundary layer in a saturated porous medium. Appl Math Lett (published online)
16. Nazar R, Amin N, Pop I (2004) Unsteady mixed convection boundary-layer flow near the stagnation point over a vertical surface in a porous medium. Int J Heat Mass Transfer 47:2681–2688
17. Pavlov KB (1974) Magnetohydrodynamic flow of an incompressible viscous fluid caused by deformation of a surface. Magnitnaya Gidro-dinamika 4:146–147
18. Pop I, Na TY (1998) A note on MHD flow over a stretching permeable surface. Mech Res Comm 25(3):263–269
19. Raptis A, Perdikis C, Takhar H S (2004) Effect of thermal radiation on MDH flow. App Math Comp 153:645–649
20. Schlichting H, Gersten K (2000) Boundary layer theory. 8th Revised and Enlarged Ed, Springer-Verlag, Berlin, Heidelberg 2000
21. Takhar HS (1971) hydromagnetic free convection from a flat plate. Indian J Phys 45:289–311
22. Takhar HS, Ali MA, Gupta AS (1989) Stability of magnetohydrodynamic flow over a stretching sheet. Liquid Metal Hydrodynamics (J. Lielpeteris and R. Moreau eds) Kluwer Academic Publishers, Dordrecht, pp 465–471
23. Takhar HS, Chamkha AJ, Nath G (2000) Flow and mass transfer on a stretching sheet with a magnetic field and chemically reactive species. Int J Eng Sci 38:1303–1314
24. Takhar HS, Nath G (1997) Similarity solutions of unsteady boundary layer equations with a magnetic field. Mecanica 32:157–163
25. Takhar HS, Raptis A, Perdikis C (1987) MHD asymmetric flow past a semi-infinite moving plate. Acta Mechanica 65:287–290
26. Vajravelu K (1986) Hydromagnetic flow and heat transfer over a continuous moving porous surface. Acta Mechanica 64:179–185
27. Watanabe T, Pop I (1995) Hall effects on magnetohydrodynamic boundary layer flow over a continuous moving flat plate. Acta Mechanica 108:35–47

Chapter 2

Computational complexity investigations for high-dimensional model representation algorithms used in multivariate interpolation problems

M.A. Tunga[1], M. Demiralp[2]

[1]Computer Engineering Department, Bahcesehir University, 34349 Besiktas, Istanbul, Turkey, alper.tunga@bahcesehir.edu.tr
[2]Informatics Institute, Istanbul Technical University, 34469 Maslak, Istanbul, Turkey, demiralp@be.itu.edu.tr

Abstract. In multivariate interpolation problems, increase in both the number of independent variables of the sought function and the number of nodes appearing in the data set causes computational and mathematical difficulties. It may be a better way to deal with less variate partitioned data sets instead of an *N*-dimensional data set in a multivariate interpolation problem. New algorithms such as high-dimensional model representation (HDMR), generalized HDMR, factorized HDMR, hybrid HDMR are developed or rearranged for these types of problems. Up to now, the efficiency of the methods in mathematical sense was discussed in several papers. In this work, the efficiency of these methods in computational sense will be discussed. This investigation will be done by using several numerical implementations.

Keywords. Data partitioning, Multivariate approximation, High-dimensional model representation, Computational complexity

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_2,

2.1 Introduction

If the values of a multivariate function $f(x_1,\ldots,x_N)$ are given for only a finite number of points in the space of its arguments and it is asked to determine an analytical structure for the sought multivariate function, standard multivariate routines may become cumbersome as the dimensionality grows. This urges us to use a divide-and-conquer algorithm which approximates the function for the mentioned multivariate interpolation problems. Hence, the given multivariate data are partitioned into low variate data and then an analytical structure determined with the aid of these partitioned data.

For this purpose, two new data partitioning methods were developed by using the philosophy given in high-dimensional model representation (HDMR) method which was first proposed by I. M. Sobol [1]. The equation given by Sobol for this method is as follows:

$$f(x_1,\ldots,x_N)=f_0+\sum_{i_1=1}^{N}f_{i_1}(x_{i_1})+\sum_{\substack{i_1,i_2=1\\ i_1<i_2}}^{N}f_{i_1,i_2}(x_{i_1},x_{i_2}) \\ +\cdots+f_{1\cdots N}(x_1,\ldots,x_N) \tag{2.1}$$

This expansion is a finite sum and is composed of a constant term, univariate terms, bivariate terms, and so on. These are the HDMR components of a given multivariate function.

Then, several other new algorithms based on this method were proposed in more comprehensive forms for different types of engineering problems by H. Rabitz, M. Demiralp, and their groups [2–11].

A multivariate function can be given by its values at a finite number of nodes of a hyperprismatic regular grid. These nodes can be represented by N tuples which are the elements of a cartesian product of the given individual sets of values for each independent variable. high-dimensional model representation is used to approximately partition these given multivariate data into low variate data [7].

On the other hand, data need not be given at all nodes of hyperprismatic regular grid. Instead, it can be given at certain randomly chosen nodes. Hence, certain level of incompleteness may be encountered in HDMR method for such data sets. This time, generalized high-dimensional model representation (GHDMR), which is based on the HDMR expansion, is used as a data partitioning technique [8].

At this point, the nature of the given data, in other words the nature of the sought function, and the construction features of the data set affect the behavior of the interpolation problem and the structure of the high-dimensional model representation method. To this end, HDMR or generalized HDMR (GHDMR) can be used to partition the multivariate data and to determine an approximate analytical structure for the sought function. These methods work well for the sought functions having additive nature as a result of the additive structure of the HDMR expansion. For the sought functions having dominantly or entirely multiplicative nature, factorized HDMR (FHDMR) is used [9,10]. Hybrid HDMR (HHDMR) method is used when the sought function has intermediate nature, that is, it has neither a dominantly additive nor a dominantly multiplicative nature [11].

These above-mentioned methods were developed and published in several journals. In this work, we will discuss CPU times spent for each algorithm in different types of multivariate interpolation problems. There exists a chapter related to the numerical testing implementations for this investigation. The results are obtained by using certain program codes (scripts) written in MuPAD 4.0, Multi Processing Algebra Data tool [12,13]. This software is developed by the MuPAD Research Group at the University of Paderborn in Germany. MuPAD is a general-purpose computer algebra system for symbolic and numerical computations. Additionally, PERL Scripting Language, Practical Extraction and Report Language, is used for making the given multivariate data amenable for MuPAD program codes [14]. MuPAD program codes run in a 20-digits precision environment. These results are obtained on a PC of P-IV 2400 MHz CPU speed and 512 MB RAM.

2.2 Data partitioning via HDMR

HDMR is constructed as an expansion for a given multivariate function such that its components are ordered starting from a constant component (zeroth-order multivariance) and continuing in ascending multivariance, that is, univariate, bivariate, trivariate components, and so on. The main step of the algorithm is to determine the right-hand-side components of the HDMR expansion given in Eq. (2.1). To obtain the structure of the constant term, the following operator is defined:

$$I_0 F(x_1,\ldots,x_N) = \int_{a_1}^{b_1} dx_1 W_1(x_1) \times\cdots\times \int_{a_N}^{b_N} dx_N W_N(x_N) F(x_1,\ldots,x_N) \tag{2.2}$$

Similarly, the following operator is defined to build a way to determine the structure of the univariate HDMR term of the given multivariate function:

$$I_m F(x_1,\ldots,x_N) = \int_{a_1}^{b_1} dx_1 W_1(x_1) \times \cdots \times \int_{a_{m-1}}^{b_{m-1}} dx_{m-1} W_{m-1}(x_{m-1}) \times \int_{a_{m+1}}^{b_{m+1}} dx_{m+1} W_{m+1}(x_{m+1}) \times \cdots \times \int_{a_N}^{b_N} dx_N W_N(x_N) F(x_1,\ldots,x_N) \tag{2.3}$$

where $1 \le m \le N$. The function $F(x_1,\ldots,x_N)$ appearing in these two relations is an arbitrary square integrable function. When the above-mentioned operators are applied to both sides of the HDMR expansion given in Eq. (2.1), the structures of the constant and univariate terms are obtained [7]. Other operators can be defined in a similar philosophy to determine the structures of the other HDMR terms, such as bivariate terms.

Additionally, to uniquely determine these components, the following vanishing conditions are used in the evaluation of the integrals appearing in the above-mentioned operators:

$$\int_{a_1}^{b_1} dx_1 \cdots \int_{a_N}^{b_N} dx_N W(x_1,\ldots,x_N) f_i(x_i) = 0 \tag{2.4}$$

where $1 \le i \le N$. Since we need to perform a multivariate interpolation on a finite number of discrete points we can extend the domain of HDMR variables to the entire space without imposing any extra conditions. Hence, we assume that the interval for each independent variable is $(-\infty, \infty)$. It is assumed that the structure of the function $f(x_1,\ldots,x_N)$ is not given analytically. Instead it is specified by values on a finite number of points of the Euclidean space defined by the independent variables $x_1,\ldots,x_N$. These points are defined through a cartesian product. For this definition, first the datum of the variable x_j is defined as the following set:

$$D_j \equiv \left\{ \xi_j^{(k_j)} \right\}_{k_j=1}^{k_j=n_j} = \left\{ \xi_j^{(1)},\ldots,\xi_j^{(n_j)} \right\} \tag{2.5}$$

where $1 \le j \le N$. The cartesian product mentioned above can be constructed from these sets as follows:

$$D \equiv D_1 \times D_2 \times \cdots \times D_N \tag{2.6}$$

The weight function appearing in the vanishing conditions is assumed to be a product of univariate functions each of which depends on a different independent variable. The structure which needs to be created through the interpolation must include the values of the function $f(x_1,\ldots,x_N)$ on the given points only. This structure can be obtained by formatting the weight function for this purpose. In this sense the necessary action is to define the weight function as a linear combination of several Dirac delta functions [15]. Hence, the following univariate weight functions are selected:

$$W_J(x_j) \equiv \sum_{k_j=1}^{n_j} \alpha_{k_j}^{(j)} \delta\left(x_j - \xi_j^{(k_j)}\right), \quad x_j \in \left[a_j, b_j\right], \quad 1 \le j \le N \tag{2.7}$$

Using this weight function the operators mentioned in this section can be applied to both sides of the HDMR expansion by the help of the vanishing conditions and the given multivariate data are partitioned into low-variate data sets. In this work we deal with constant, univariate, and at most bivariate terms.

After several integrations a constant value, univariate partitioned data set, and bivariate partitioned data set are obtained. To this end, we have a constant value, n_m ordered pairs for the univariate function $f_m(x_m)$, and $n_{m_1} n_{m_2}$ ordered pairs for the bivariate function $f_{m_1 m_2}(x_{m_1}, \backslash x_{m_2})$ [7]. The next step is to determine analytical structures for these partitioned data sets and build the HDMR expansion of the sought function by using these structures. This step will be given in the fourth section of this chapter. The next section is about another data partitioning technique.

2.3 Data partitioning via GHDMR

If a multivariate datum is given for the determination of a multivariate function, the location of data points in hyperspace of the independent variables gains a lot of importance. If they are located at the points of a set which is constructed as a direct product of univariate sets, high-dimensional model representation (HDMR) can be successfully used to partition the data into less variate data. On the other hand HDMR becomes unemployable when the data are random or not given at all points of a grid which is constructed via direct product of univariate meshes due to the incompleteness of the data. Hence, for these cases, a new high-dimensional model representation method is needed. generalized high-dimensional model representation (GHDMR) is used for this purpose. In this method a general multivariate weight function is used instead of a product-type weight function. The algorithm uses the HDMR components of this general weight function. The steps of the method include first the determination of the HDMR components of the general multivariate weight function by using a product-type auxiliary weight function.

$$\Omega(x_1,\ldots,x_N)\equiv\prod_{j=1}^{N}\Omega_j(x_j) \tag{2.8}$$

Then, these components are employed in the formulae to obtain the GHDMR components of the given multivariate function. In this way, the multivariate data are partitioned into low-variate data. Here, the constant and the univariate terms of GHDMR expansion are obtained to get an approximation. Similar operators as given in the first section are used for this purpose. This time, the integrations will be evaluated by also using the HDMR components of the general weight function under the auxiliary weight function. The following orthogonality conditions are employed in these evaluations:

$$\int_{a_1}^{b_1}dx_1\times\cdots\times\int_{a_N}^{b_N}dx_N\Omega(x_1,\ldots,x_N)W(x_1,\ldots,x_N)f_i(x_i)=0 \tag{2.9}$$

where$1\le i\le N$. As a result constant and univariate GHDMR terms are obtained. Relation for the univariate terms corresponds to an integral equation system whose unknowns are the univariate GHDMR terms [8].

When we use this method to partition the multivariate random data, the following general weight function is selected:

$$W(x_1,\ldots,x_N)\equiv\sum_{j=1}^{m}\alpha_j\delta\left(x_1-x_1^{(j)}\right)\times\cdots\times\delta\left(x_N-x_N^{(j)}\right) \tag{2.10}$$

where α_j parameters are used for making it possible to give different importance to each individual datum.

Using this general weight function and the orthogonality conditions given in Eq. (2.9) when applying the above-mentioned operators to the HDMR expansion, a constant value and a number of linear equations whose unknowns are the univariate component values at the given data points of *N*-dimensional space are obtained. The final step of this algorithm is to determine the unknowns of this equation set. This completes the construction of the univariate components at the data points [8].

At this point, approximate analytical structure should be determined by using these partitioned data. For this purpose, Lagrange interpolation formula will be used. The next section is about this subject.

2.4 Interpolation

By partitioning the given multivariate data via HDMR or GHDMR a table of pairs of data can be obtained instead of an analytical structure for the function $f_m(x_m)$. This table provides an opportunity to determine the function $f_m(x_m)$ under an assumed structure, that is, to interpolate the corresponding data. By this way, multivariate interpolation, at least for these functions, can be approximately reduced to a set of univariate interpolations. To determine the overall structure of the function, an analytical structure should be defined or a calculation rule should be imposed on the interpolation. If the function to be determined by HDMR or GHDMR is sufficiently smooth, then the function can be represented with a multinomial of all independent variables over the continuous region produced by the cartesian product of the related intervals. For this purpose, first a multinomial representation should be built for $f_m(x_m)$:

$$p_m(x_m)=\sum_{k_m}^{n_m} L_{k_m}(x_m) f_m(\xi_m^{(k_m)}), \quad \xi_m^{(k_m)} \in D_m, \; 1 \le m \le N \tag{2.11}$$

Here $L_{k_m}(x_m)$s are Lagrange coefficient polynomials [16] which are independent of the structure of the function. The structures of these polynomials are given:

$$L_{k_m}(x_m) \equiv \prod_{\substack{j=1 \\ j \ne k_m}}^{n_m} \frac{\left(x_m-\xi_m^{(j)}\right)}{\left(\xi_m^{(k_m)}-\xi_m^{(j)}\right)}, \quad \xi_m^{(k_m)} \in D_m, \; 1 \le k_m \le n_m, \; 1 \le m \le N \tag{2.12}$$

As Lagrange polynomials are constructed, univariate functions given by the relation Eq. (2.11) are uniquely determined within continuous polynomial interpolation. These functions can be considered as univariate components of HDMR or GHDMR for the multivariate function $f(x_1,\ldots,x_N)$. The expansion formed by the summation of these functions and the constant term provides the following multinomial approximation which is called “univariate approximation”:

$$s_1(x_1,\ldots,x_N)=f_0+\sum_{m=1}^{N}p_m(x_m) \tag{2.13}$$

Same relations for the higher variate approximations can be defined in a similar way.

2.5 Factorized HDMR

We have observed that the truncations of both HDMR and GHDMR work well as long as the multivariate function under consideration has additive nature. If it is completely additive then data partitioning is exact, otherwise a certain level of truncation error is encountered. Additivity is one end of the behavior of the multivariate function. The other hand is multiplicativity where all HDMR components contribute to the function at similar orders. Therefore, truncation approximation fails to describe the multivariate function under consideration. In those cases we need to formulate a different truncation approximation which somehow takes all components of HDMR or GHDMR into consideration. The first step is to write this new equality (FHDMR) for this method:

$$\begin{aligned} f(x_1,\ldots,x_N)=r_0&\left[\prod_{i_1=1}^{N}\left(1+r_{i_1}(x_{i_1})\right)\right]\left[\prod_{\substack{i_1,i_2=1\\ i_1<i_2}}^{N}\left(1+r_{i_1i_2}(x_{i_1},x_{i_2})\right)\right] \\ &\times\cdots\times\left[\left(1+r_{i_1\cdots N}(x_1,\ldots,x_N)\right)\right] \end{aligned} \tag{2.14}$$

The right-hand-side components of the above relation can be determined by making comparisons between the right-hand-side of equation Eq. (2.1) and the additive form of the right-hand-side in Eq. (2.14). To make comparisons, idempotent operators will be used as auxiliary tools. These operators satisfy the following relations:

$$I_j^{(id)}I_k^{(id)}\equiv I_k^{(id)}I_j^{(id)}, \qquad \left[I_j^{(id)}\right]^2\equiv I_j^{(id)} \tag{2.15}$$

where $j,k=1,\ldots,N$. Using these operators HDMR and FHDMR expansions are replaced by the following generalized ones:

$$S(x_1,\ldots,x_N)=f_0I+\sum_{i_1=1}^{N}f_{i_1}(x_{i_1})I_{i_1}^{(id)}+\cdots \tag{2.16}$$

$$R(x_1,\ldots,x_N)=r_0\left[\prod_{i_1=1}^{N}\left(I+r_{i_1}(x_{i_1})I_{i_1}^{(id)}\right)\right]\times\cdots$$

These two entities represent the same multivariate function. Hence, their right-hand-sides must match for all idempotent operators. This permits us to determine the constant term, the univariate terms, and higher order terms of the FHDMR expansion.

As a result, constant, univariate, and bivariate FHDMR terms are obtained in terms of HDMR or GHDMR as follows:

$$r_0=f_0 \tag{2.17}$$

$$r_{i_1}(x_{i_1})=\frac{f_{i_1}(x_{i_1})}{f_0}$$

$$r_{i_1i_2}(x_{i_1},x_{i_2})=\frac{f_0f_{i_1i_2}(x_{i_1},x_{i_2})-f_{i_1}(x_{i_1})f_{i_2}(x_{i_2})}{\left(f_0+f_{i_1}(x_{i_1})\right)\left(f_0+f_{i_2}(x_{i_2})\right)}$$

2.6 Hybrid HDMR

In most cases the given multivariate data and the sought multivariate function have neither purely additive nor purely multiplicative nature. They have a hybrid nature. So, a new method is used to obtain better results and it is called hybrid high-dimensional model representation (HHDMR). This new expansion includes both the HDMR (or GHDMR) and the FHDMR expansions through a hybridity parameter, γ:

$$f(x_1,\ldots,x_N)=\gamma\left(f_0+\sum_{i_1=1}^{N}f_{i_1}(x_{i_1})+\cdots\right) \tag{2.18}$$

$$+(1-\gamma)\left(r_0\left[\prod_{i_1=1}^{N}\left(1+r_{i_1}(x_{i_1})\right)\right]\times\cdots\right)$$

Using Eq. (2.18) an HHDMR approximant can be defined as follows by using the HDMR and the FHDMR approximants:

$$h_{jk}(x_1,\ldots,x_N;\gamma)=\gamma s_j(x_1,\ldots,x_N)+(1-\gamma)\pi_k(x_1,\ldots,x_N),\qquad(2.19)$$
$$0\le j,k\le N$$

where $s_j(x_1,\ldots,x_N)$ stands for the jth HDMR approximant and $\pi_k(x_1,\ldots,x_N)$ stands for the kth FHDMR approximant which is a truncated product including at most k-variate factors.

The most important step here is to determine the hybridity parameter γ. For this purpose, a functional is defined as

$$F(\gamma)\equiv\left\|f_{\text{org}}-f_{\text{HHDMR}}(\gamma)\right\|^2\qquad(2.20)$$

where f_{org} and f_{HHDMR} stand for the original function and the function obtained from the HHDMR expansion respectively. We need to obtain the γ value that minimizes the value of this norm. This minimization criterion can be written as

$$\frac{\partial F}{\partial\gamma}=0\qquad(2.21)$$

Using this criterion, the best value for that parameter can be obtained [11]. By this way the best representation for the sought multivariate function can be determined via hybrid HDMR.

2.7 Error analysis

According to the above-mentioned methods, HDMR or GHDMR, FHDMR, and HHDMR, several representations can be obtained approximately by using the constant, univariate, and bivariate terms of the mentioned expansions. For obtaining these several representations there exist questions, that is, how to find the best expansion for the sought multivariate function or whether the obtained representations are or are not the acceptable solutions for the given engineering problems. For this purpose, the following relative norm

$$N=\frac{\left\|f_{org}-f_{new}\right\|}{\left\|f_{org}\right\|}\qquad(2.22)$$

will be evaluated. Here, f_{new} stands for the multivariate function obtained via a high-dimensional model representation expansion.

The minimum norm value obtained by using this relation through all the evaluated norm values will show the best representation for the sought multivariate function. This result is assumed to be the best representation for the multivariate function.

2.8 Numerical implementations

In this section, the numerical implementations are classified into two main parts. The first part includes the examples in which the HDMR method is used as a data partitioning technique. In this part FHDMR and HHDMR algorithms are used to partition data obtained through HDMR. In the second part, the examples are constructed by using GHDMR method. FHDMR and HHDMR algorithms are used to partition data which are obtained through the GHDMR method.

The results are obtained by using MuPAD 4.0. The CPU time results for each implementation are evaluated by using "time()" function which returns the total CPU time in milliseconds that was spent by the current MuPAD process. Only the relative error values and CPU times spent for the evaluations are given in this work.

It is assumed that the following *(N+1)* – tuples are taken as data to describe a multivariate function $f(x_1,\ldots,x_N)$

$$d_j \equiv \left(x_1^{(j)},\ldots,x_N^{(j)},\varphi_j\right), \qquad 1 \le j \le m \tag{2.23}$$

where φ_j is the value of $f(x_1,\ldots,x_N)$, the sought function, at the point described by the first N components of d_j in the N-dimensional space we are concerned. That is,

$$\varphi_j \equiv f\left(x_1^{(j)},\ldots,x_N^{(j)}\right), \qquad 1 \le j \le m \tag{2.24}$$

To construct the information for the data set which is the values of the sought multivariate function at the nodes of the grid, analytical structures of known multivariate functions are used.

2.9 HDMR-based implementations

The first example considered here is a multivariate function which is completely additive, that is, the sum of univariate functions as follows:

$$f(x_1,\ldots,x_{10})=\sum_{i=1}^{10} a_i x_i, \qquad a_i = 2i-1 \tag{2.25}$$

This function has 10 independent variables and it is assumed that the given data set has 16,384 nodes in it. The relative error value obtained for the univariate HDMR approximant and the CPU time spent for this approximation are

$$N_{s_1} = 2.84\times10^{-25}, \qquad t_{s_1} = 4.48 \text{ mins} \tag{2.26}$$

respectively. Because the programming environment has 20 decimal digit accuracy, this result can be assumed to be zero and it means that the representation obtained is exact for the multivariate function dealt with.

In the second example, the selected multivariate function has five independent variables where the function is of purely multiplicative nature

$$f(x_1,x_2,x_3,x_4,x_5)=x_1 x_2 x_3 x_4 x_5 \tag{2.27}$$

and there are 640 nodes in the given hyperprismatic regular grid. The results of the relative error analysis and the CPU times spent for each algorithm are obtained as follows:

$$\begin{aligned} N_{s_1} &= 3.16\times10^{-1}, & t_{s_1} &= 1.56 \text{ secs} \\ N_{s_2} &= 8.66\times10^{-2}, & t_{s_2} &= 6.44 \text{ secs} \\ N_{\pi_1} &= 8.37\times10^{-25}, & t_{\pi_1} &= 6.47 \text{ secs} \end{aligned} \tag{2.28}$$

where s_1, s_2, and π_1 correspond to the univariate HDMR, bivariate HDMR, and univariate FHDMR approximants.

The analytical structure of the multivariate function is defined as follows as the third example with six independent variables:

$$f(x_1,x_2,x_3,x_4,x_5,x_6)=(x_1+x_2+x_3+x_4+x_5+x_6)^5 \tag{2.29}$$

In this example the given data set is constructed by using 6400 nodes. The relative error values and the CPU times are obtained as follows:

$$\begin{aligned}
N_{s_1} &= 7.30\times10^{-2}, & t_{s_1} &= 9.32\,\text{secs} \\
N_{s_2} &= 7.43\times10^{-3}, & t_{s_2} &= 22.79\,\text{secs} \\
N_{\pi_1} &= 1.92\times10^{-2}, & t_{\pi_1} &= 9.34\,\text{secs} \\
N_{\pi_2} &= 1.14\times10^{-3}, & t_{\pi_2} &= 22.93\,\text{secs} \\
N_{h_{11}} &= 5.13\times10^{-3}, & t_{h_{11}} &= 19.37\,\text{secs} \\
N_{h_{22}} &= 6.06\times10^{-4}, & t_{h_{22}} &= 391.15\,\text{secs}
\end{aligned} \tag{2.30}$$

2.10 GHDMR-based implementations

In the following example we know the nodes of the mesh and the values of the sought function at the nodes of the given mesh. Hence, the domains for the independent variables are known. For the following numerical implementation there are 4,976,640 nodes in the mesh. From this mesh 100 nodes are selected randomly. Using these nodes and the values of the following selected multivariate function at these nodes a multivariate data set is constructed:

$$f(x_1,\dots,x_{10}) = \sum_{i=1}^{10} i x_i \tag{2.31}$$

The relative error value and the CPU time spent for this generalized form HDMR method are obtained as follows:

$$N_{\bar{s}_1} = 1.76\times10^{-25}, \qquad t_{\bar{s}_1} = 7.69\,\text{secs} \tag{2.32}$$

where $\bar{s}_1$ corresponds to the univariate GHDMR approximant.

The last example is given to discuss the performance results of GHDMR, FHDMR, and HHDMR methods for the following multivariate interpolation problem. The analytical structure of the sought function is selected as

$$f(x_1,x_2,x_3,x_4,x_5) = \prod_{i=1}^{5} (1+4x_i) \tag{2.33}$$

where the problem has 100 nodes. It has both additive and multiplicative features. Hence, it is expected that the HHDMR approximants will give better results than GHDMR and FHDMR. To make this comparison the following relative error values of all approximants obtained through

GHDMR, FHDMR, and HHDMR are calculated, and needed CPU times for these calculations are measured:

$$\begin{aligned} N_{\bar{s}_1} &= 1.84\times 10^{-1}, & t_{\bar{s}_1} &= 2.10\,\text{secs} \\ N_{\pi_1} &= 1.03\times 10^{-1}, & t_{\pi_1} &= 2.16\,\text{secs} \\ N_{h_{11}} &= 7.52\times 10^{-2}, & t_{h_{11}} &= 8.25\,\text{secs} \end{aligned} \tag{2.34}$$

2.11 Concluding remarks

In this work, the basic idea is to partition the given data to less variate data and then to interpolate them individually to fit an analytical structure to the multivariate function to be determined. The elements of data set are assumed to be given at the nodes of a hyperprismatic grid. Certain nodes may be missing to locate data or entire nodes are used to specify the values of the multivariate function under consideration. If data are given at all nodes of a hyperprismatic grid then classical HDMR can be used for partitioning. On the other hand, GHDMR should be used instead of HDMR when the data have no datum for certain nodes. The nature of the sought multivariate function also affects the method in use. Since the HDMR expansion has an additive structure, these two methods seem to be effective for additive-type functions. As the sought function has not only additive but also multiplicative nature, the obtained representation via HDMR or GHDMR for the sought function gets worse. Hence, certain new methods are needed to determine better representations for the functions having multiplicative or intermediate natures. For this purpose, FHDMR and HHDMR methods are used.

As a result, we have HDMR, GHDMR, FHDMR, and HHDMR methods to deal with the functions whose nature is additive or multiplicative or intermediate type.

When the results given in the previous section are examined carefully depending on the nature of the sought multivariate function the results get better while we use the method that best fits. However, when the number of nodes or the number of HDMR terms taken into consideration increases, more time periods are needed to obtain better results. This brings much more CPU time need for the mentioned algorithms. This means that if you want the best solution for your problem you have to wait much more for the results. On the other hand, if a result obtained by using an approximant having less variate terms is sufficient for the given problem, then you may spend less CPU time for your work.

Acknowledgment. The second author is grateful to Turkish Academy of Sciences and both authors thank WSEAS for their supports.

References

1. Sobol IM (1993) Sensitivity estimates for nonlinear mathematical models. Math. Model. and Comput. Exp. (MMCE) 1:407
2. Rabitz H, Alış Ö (1999) General foundations of high dimensional model representations. J. Math. Chem. 25:197–233
3. Alış Ö, Rabitz H (2001) Efficient implementation of high dimensional model representations. J. Math. Chem. 29:127–142
4. Li G, Rosenthal C, Rabitz H (2001) High dimensional model representations. J. Math. Chem. A 105:7765–7777
5. Demiralp M (2003) High dimensional model representation and its varieties. Math. Res. 9:146–159
6. Tunga B, Demiralp M (2003) Hybrid high dimensional model representation approximants and their utilization in applications. Math. Res. 9:438–446
7. Demiralp M, Tunga MA (2001) High dimensional model representation of multivariate interpolation via hypergrids. In: the Sixteenth International Symposium on Computer and Information Sciences, pp 416–423
8. Tunga MA, Demiralp M (2003) Data partitioning via generalized high dimensional model representation (GHDMR) and multivariate interpolative applications. Math. Res. 9:447–462
9. Tunga MA, Demiralp M (2004) A factorized high dimensional model representation on the partitioned random discrete data. Appl. Num. Anal. Comp. Math. 1:231–241
10. Tunga MA, Demiralp M (2005) A factorized high dimensional model representation on the nodes of a finite hyperprismatic regular grid. Appl. Math. and Comput. 164:865–883
11. Tunga MA, Demiralp M (2006) Hybrid high dimensional model representation (HHDMR) on the partitioned data. J. Comput. Appl. Math. 185:107–132
12. Oevel W, Postel F, Wehmeier S, Gerhard J (2000) The MuPAD tutorial. Springer, New York
13. http://www.mupad.de
14. Deitel HM, Deitel PJ, Nieto TR, McPhie DC (2001) Perl how to program. Prentice Hall, Englewood Cliffs
15. Zemanian AH (1987) Distribution theory and transform analysis, An introduction to generalized functions, with applications. Dover Publications Inc., New York
16. Buchanan JL, Turner PR (1992) Numerical methods and analysis. McGraw-Hill, New York

Chapter 3

On competition between modes of the onset of Marangoni convection with free-slip bottom under magnetic field

N.M. Arifin, H. Rosali

Department of Mathematics, Faculty of Science, Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia, norihan@math.upm.edu.my

Abstract. In this chapter we use a numerical technique to analyze the onset of Marangoni convection in a horizontal layer of electrically conducting fluid heated from below and cooled from above in the presence of a uniform vertical magnetic field. The top surface of a fluid is deformably free and the bottom boundary is rigid and free-slip. The critical values of the Marangoni numbers for the onset of Marangoni convection are calculated and the latter is found to be critically dependent on the Hartmann, Crispation, and Bond numbers. In particular, we present an example of a situation in which there is competition between modes at the onset of convection.

Keywords. Marangoni convection, Magnetic field, Free-slip

3.1 Introduction

Convection in a plane horizontal fluid layer heated from below, known as the Rayleigh–Benard convection, is the type of convection considered most frequently. Rayleigh [6] was the first to solve the problem of the onset of thermal convection in a horizontal layer of fluid heated from below. His linear analysis showed that Benard convection occurs when the Rayleigh number exceeds a critical value. Theoretical analysis of Marangoni convection was started with the linear analysis by Pearson [5] who

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_3,

assumed an infinite fluid layer, a nondeformable case, and zero gravity in the case of no-slip boundary conditions at the bottom. He showed that thermocapillary forces can cause convection when the Marangoni number exceeds a critical value in the absence of buoyancy forces. Pearson [5] obtained the critical Marangoni number M_c = 79.607 and the critical wave number a_c = 1.9929. Linear stability analysis of Marangoni convection with free-slip boundary conditions at the bottom was first investigated by Boeck [1]. For free-slip case, Boeck [1] obtained the critical Marangoni number M_c = 57.598 and the critical wave number a_c = 1.7003.

The effect of a magnetic field on the onset of steady buoyancy and thermocapillary-driven (Benard–Marangoni) convection in a fluid layer with a nondeformable free surface was first analyzed by Nield [4]. He found that the critical Marangoni number monotonically increased as the strength of vertical magnetic field increased. This indicates that Lorentz force suppressed Marangoni convection. Later, the effect of a magnetic field on the onset of steady Marangoni convection in a horizontal layer of fluid has been discussed in a series of paper by Wilson [7–9]. The influence of a uniform vertical magnetic filed on the onset of oscillatory Marangoni convection was treated by Hashim and Wilson [3] and Hashim and Arifin [2].

The above investigators pertain their analyses to Marangoni convection in the presence of magnetic field with no-slip lower boundary condition. In this study, we consider the onset of steady Marangoni convective instability in a horizontal fluid layer of electrically conducting fluid with a deformable upper free surface and a free-slip lower surface, subject to a uniform magnetic field. To the author's best knowledge this problem has not been reported in the literature. The linear stability theory is applied and the resulting eigenvalue problem is solved numerically. The effects of the Hartmann number and a free surface deformation on the onset of steady Marangoni convection are studied.

3.2 Problem formulation

Consider a horizontal fluid layer of depth d heated from below, subject to a uniform vertical magnetic field and a uniform vertical temperature gradient. The fluid layer is bounded below by a horizontal solid boundary at constant temperature T_1 and above by a free surface at constant temperature T_2 which is in contact with passive gas at pressure P_o and constant temperature T_∞.

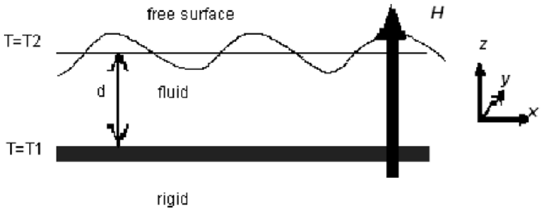


Fig. 3.1. Geometry of the unperturbed state

We used cartesian coordinates with two horizontal *x*- and *y*-axes which are located at the lower solid boundary and a positive *z*-axis that is directed toward the free surface. The surface tension τ is assumed to be a linear function of the temperature:

$$\tau = \tau_o - \gamma(T - T_o) \tag{3.1}$$

where τ_o is the value of τ at temperature T_o and the constant γ is positive for most fluids. The density of the fluids is given by

$$\rho = \rho_o \{1 - \alpha(T - T_o)\} \tag{3.2}$$

where α is the positive coefficient of the thermal liquid expansion and ρ_o is the value at the reference temperature T_o.

Subject to the Boussinesq approximation, the governing equations for an incompressible, electrically conducting fluid in the presence of a magnetic field are

$$\nabla.\mathbf{U}=0 \tag{3.3}$$

$$\nabla.\mathbf{H}=0 \tag{3.4}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{U}.\nabla\right)\mathbf{U} = -\frac{1}{\rho}\nabla\Pi + \nu\nabla^2\mathbf{U} + \frac{\mu}{4\pi\rho}(\mathbf{H}.\nabla)\mathbf{H} \tag{3.5}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{U}.\nabla\right)\mathbf{H} = (\mathbf{H}.\nabla)\mathbf{U} + \eta\nabla^2\mathbf{H} \tag{3.6}$$

$$\left(\frac{\partial}{\partial t} + \mathbf{U}.\nabla\right)T = \kappa\nabla^2 T \tag{3.7}$$

where **U** is the fluid velocity, **H** is the magnetic field, *T* is the temperature, ν is the kinematic viscosity, κ is the thermal diffusivity, η is the electrical resistivity, and $\Pi = p + \mu|\mathbf{H}|^2/8\pi$ is the magnetic pressure, where p is the fluid pressure and μ is the magnetic permeability. When motion occurs the upper free surface of the layer will be deformable with its position at $z = d + f(x,y,t)$. At the free surface, we have the usual kinematic condition together with conditions of continuity of the normal and tangential stresses, and the temperature obeys Newton's law of cooling, $k\partial T/\partial\mathbf{n} = h(T - T_\infty)$, where k and h are the thermal conductivity of the fluid and the heat transfer coefficient between the free surface and the air, respectively, and $\mathbf{n}$ is the outward unit normal to the free surface. At the lower rigid boundary the usual no-slip conditions requires continuity of velocity between the solid and the fluid.

To simplify the analysis, it is convenient to write the governing equations and boundary conditions in a dimensionless form. In the dimensionless formulation, scales for length, time, velocity, temperature, and magnetic field have been taken to be $d, d^2/\nu, \nu/d, \beta d\nu/\kappa$, and $\mu\overline{H}/\eta$, respectively, where $\overline{H}$ is the initial magnetic field strength. Furthermore, six dimensionless groups appearing in the problem are the Marangoni number $M = \gamma\beta d^2/\rho\nu\kappa$, the Hartmann number (the square root of the Chandrasekhar number) $H = \mu\overline{H}d(\sigma/\rho\nu)^{1/2}$, the Biot number $B_i = hd/k$, the Bond number $B_o = \rho g d^2/\tau_o$, the Prandtl number $P_1 = \nu/\kappa$ and the magnetic Prandtl number $P_2 = \nu/\eta$, and the Crispation number $C_r = \rho\nu\kappa/\tau_0 d$.

3.3 Linearized problem

The linearized equations and boundary conditions governing the onset of Marangoni convection in an initially quiescent horizontal fluid layer bounded above by a deformable free surface and bounded below by a thermally conducting planar boundary subject to a uniform vertical magnetic field and a uniform temperature gradient have been obtained by several authors (see, for example, Hashim and Ariffin [2]) and are given by

$$(D^2 - a^2)T + w = 0 \tag{3.8}$$

$$\left[(D^2 - a^2)^2 - H^2D^2\right]w = 0 \tag{3.9}$$

subject to

$$w = 0 \tag{3.10}$$

$$P_1 C_r \left[\left(D^2 - 3a^2 - H^2 - s \right) Dw \right] - a^2 (a^2 + B_o) f = 0 \tag{3.11}$$

$$P_1 (D^2 + a^2) w + a^2 M (P_1 T - f) = 0 \tag{3.12}$$

$$h_z = 0 \tag{3.13}$$

$$P_1 DT + B_i (P_1 T - f) = 0 \tag{3.14}$$

evaluated on the undisturbed position of the upper free surface $z = 1$, and

$$w = 0 \tag{3.15}$$

$$D^2 w = 0 \tag{3.16}$$

$$h_z = 0 \tag{3.17}$$

$$T = 0 \tag{3.18}$$

on $z = 0$ where the condition of free-slip corresponds to Eq. (3.16). The operator $D = \left(d/dz \right)$ denotes differentiation with respect to the vertical coordinate z. The variables w, T, h_z, and f denote, respectively, the vertical variation of the z-velocity, temperature, magnetic field, and the magnitude of the free surface deflection of the linear perturbation to the basic state with total wave number a in the horizontal x–y plane and complex growth rate s.

3.4 Solution of the linearized problem

In the general case $s = 0$, we follow the solution approach of Hashim and Wilson [3] and seek asymptotic solutions for w, T in the form

$$w(z) = ACe^{\xi z}, \quad T(z) = Ce^{\xi z} \tag{3.19}$$

where the exponent ξ and the complex constants A and C are to be determined. Substituting these forms into Eqs. (3.8) and (3.9) and eliminating A and C we obtain a sixth-order algebraic equation for ξ, namely

$$(\xi^2 - a^2)\left[(\xi^2 - a^2 - s)^2 - H^2\xi^2\right] = 0 \tag{3.20}$$

with six distinct roots, which we denote by $\xi_1,...,\xi_6$. Where the values of $\xi_1,...,\xi_4$ are solutions of the fourth-order algebraic equation

$$(\xi^2 - a^2 - s)^2 - H^2\xi^2 = 0 \tag{3.21}$$

while $\xi_5 = a$ and $\xi_6 = -a$. Denoting the values of A and C corresponding to ξ for $i = 1,...,6$ by A_i and C_i, respectively, we can use Eq. (3.9) to determine A_i. We can use Eq. (3.11) to eliminate the free surface deflection

$$f = \frac{P_1 C_r (D^2 - 3a^2 - H^2)Dw}{a^2(a^2 + B_o)} \tag{3.22}$$

evaluated on $z = 1$, leaving the six boundary conditions Eqs. (3.10), (3.12), (3.14), (3.15), (3.16), and Eq. (3.18) to determine the six unknowns $C_1,...,C_6$, and the general solution to the stability problem therefore

$$w(z) = \sum_{j=1}^{6} A_j C_j e^{\xi_j z}\ ,\ T(z) = \sum_{j=1}^{6} C_j e^{\xi_j z} \tag{3.23}$$

The dispersion relation between M, a, C_r, H^2, B_o, and B_i is determined by substituting these solutions into boundary conditions and evaluating the resulting 6×6 real determinants of the coefficients of the unknowns, which can be written in the form $M = -D_1/D_2$, where the two 6×6 real determinants D_1 and D_2 are independent of M.

After some simplification the elements of the determinant $D_1 = |d_{ij}|$ are given by

$$d_{1i} = A_i e^{\xi_i} \tag{3.24}$$

$$d_{2i} = \xi_i^2 A_i e^{\xi_i} \tag{3.25}$$

$$d_{3i} = (\xi_i + B_i)e^{\xi_i} \tag{3.26}$$

$$d_{4i} = A_i \tag{3.27}$$

$$d_{5i} = \xi_i^2 A_i \tag{3.28}$$

$$d_{6i} = 1 \tag{3.29}$$

for $i = 1,\dots,6$. The coefficients of the determinant D_2 are the same as those of D_1 apart from the terms

$$d_{2i} = a^2 \left[1 - \frac{C_r(\xi_i^2 - 3a^2 - H^2)\xi_i A_i}{a^2(a^2 + B_o)} \right] e^{\xi_i} \tag{3.30}$$

$$d_{3i} = \xi_i^2 e^{\xi_i} \tag{3.31}$$

for $i = 1,\dots,6$. Notice that D_1 is independent of C_r and B_o and that D_2 is independent of B_i. We could express D_1 and D_2, and hence *M*, explicitly in terms of hyperbolic functions, but since its value must then be evaluated numerically we gain little over direct numerical evaluation.

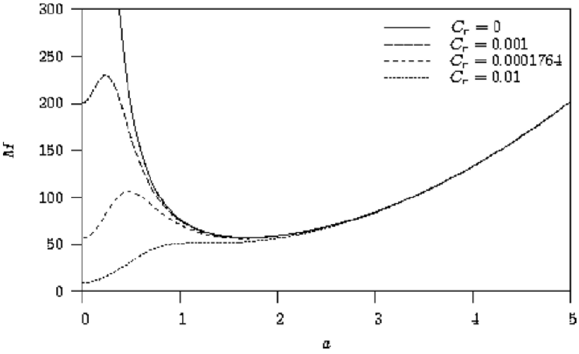


Fig. 3.2. Numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for various values of Crispation numbers, C_r, in the case $H = 0$, $B_i = 0$, and $B_o = 0.1$

3.5 Results

The effect of a magnetic field on the onset of Marangoni convection in a fluid layer with free-slip bottom in the case of a deformable free surface $(C_r \neq 0)$ is investigated numerically. Before presenting the numerical results, it is helpful to specify the range for parameters B_i, B_o, and C_r which are, respectively, given by $10^{-3} \le B_i \le 10^{-1}$, $10^{-3} \le B_o \le 10^{-1}$, and $10^{-6} \le C_r \le 10^{-2}$ for most fluid layers of depths ranging from 0.01 to 0.1 cm and are in contact with air. All numerical calculations reported in this chapter are done for the case $B_i = 0, B_o = 0.1$, and $P_1 = 1$.

Figure 3.2 shows the numerically calculated steady marginal stability curves plotted for different values of Crispation number C_r in the case $H = 0$. The Crispation number C_r, associated with the inverse effect of the surface tension, represents the degree of the free surface deformability. When C_r becomes large (corresponding to weak surface tension), the marginal curve has global minimum at zero wave number. In contrast, for small values of C_r, the marginal curve has global minimum at zero wavenumber. At some transition value of C_r, the marginal curve has two local minima, that is, one at zero wave number and the other at nonzero wave number. The transition value of C_r for the case shown in Fig. 3.2 is $C_r \approx 0.0001764$. For C_r greater than 0.0001764, the wave number at marginal stability suddenly drops from nonzero number to zero. Similar competition between different modes was identified by Hashim and Arifin [2] in the case of no-slip condition.

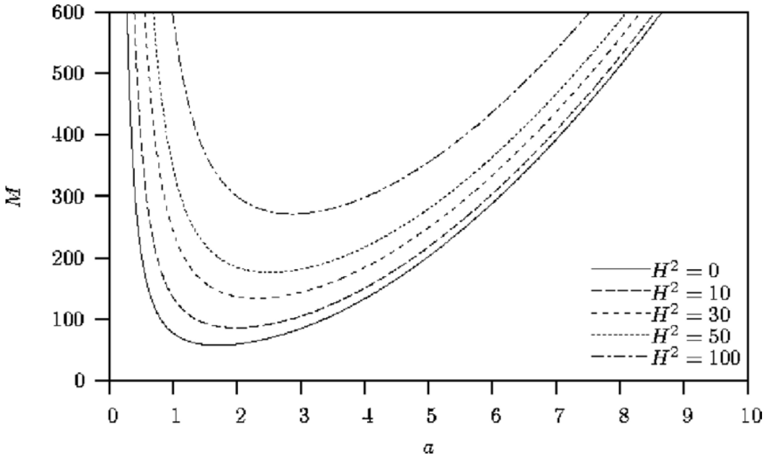


Fig. 3.3. Numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for various values of Hartmann numbers, *H*, in the case $C_r = 0$, $B_i = 0$, and $B_o = 0.1$

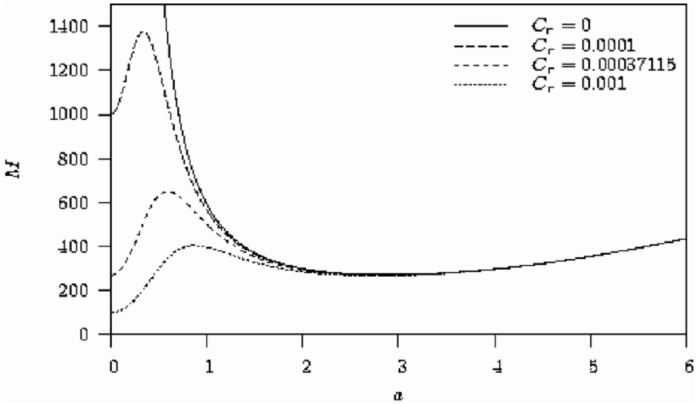


Fig. 3.4. Numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for various values of Crispation numbers, C_r, in the case $H^2 = 100$, $B_i = 0$, and $B_o = 0.1$

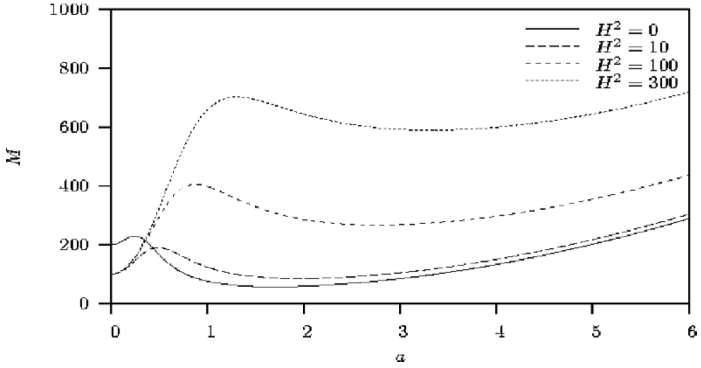


Fig. 3.5. Numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for various values of Hartmann numbers, *H*, in the case $C_r = 0.001$, $B_i = 0$, and $B_o = 0.1$

Figure 3.3 shows the numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for different values of the Hartmann number, *H*, in the case $C_r = 0$. From Fig. 3.3 it is seen that the critical Marangoni number increases with an increase of the Hartmann number. Thus, the magnetic field always has a stabilizing effect on the flow. Numerically calculated Marangoni number, *M*, as a function of the wave number, *a*, for different values of $C_r \neq 0$ in the case $H^2 = 100$ is shown in Fig. 3.4.

The figure shows parts of the marginal stability curves in the case $C_r = 0.00037115$ and $H^2 = 100$ in which zero mode (infinite wavelength) and nonzero mode (finite wavelength) occur simultaneously at the onset of convection.

Figure 3.5 shows the numerically calculated Marangoni number, *M*, as a function of the wavenumber, *a*, for different values of the Hartmann number, *H*, in the case $C_r = 0.001$. In this case, the marginal stability curve has a global minimum at the nonzero value of *a* without a magnetic field. But, the marginal stability curve always has a global minimum at zero value in the limit of a large magnetic field. We also found that two steady modes occur simultaneously at the onset of convection when $H^2 = 10$.

3.6 Conclusions

The effect of magnetic field on the onset of steady Marangoni convection in a horizontal layer of electrically conducting fluid which is free above and rigid below with free-slip condition has been studied. If the free surface is nondeformable, the absence of a magnetic field always has the stabilizing effect of increasing the critical Marangoni number for the onset of steady convection. If the free surface is deformable, then all the marginal stability curves have two local minima. The linear analysis presented in this work revealed a situation in which two steady modes compete at the onset of convection.

References

1. Boeck T, Thess A (1997) Inertial Benard-Marangoni convection. J Fluid Mech 350:149–175
2. Hashim I, Arifin NM (2003) Oscillatory Marangoni convection in a conducting fluid layer with a deformable free surface in the presence of a vertical magnetic field. Acta Mech 164:199–215
3. Hashim I, Wilson SK (1999) The effect of a uniform vertical magnetic field on the onset of oscillatory Marangoni convection in a horizontal layer of conducting fluid. Acta Mech 132:129–146
4. Nield DA (1966) Surface tension and Buoyancy effects in cellular convection of an electrically conducting liquid in a magnetic field. Z Angew Math Mech 17:131–139
5. Pearson JRA (1958) On convection cells induce by surface tension. J Fluid Mech 4:489–500

6. Rayleigh R (1916) On convection currents in a horizontal layer of fluid, when the higher temperature is on the under side. Phil Mag 32:529–546
7. Wilson SK (1993) The effect of a uniform magnetic field on the onset of steady Benard-Marangoni convection in a layer of conducting fluid. J Engng Math 27:161–188
8. Wilson SK (1993) The effect of a uniform magnetic field on the onset of Marangoni convection in a layer of conducting fluid. Q J Mech Appl Math 46:211–248
9. Wilson SK (1994) The effect of a uniform magnetic field on the onset of steady Marangoni convection in a layer of conducting fluid with a prescribed heat flux at its lower boundary. Phys Fluid 6:3591–3600

Chapter 4

Mathematical modeling of boundary layer flow over a moving thin needle with variable heat flux

S. Ahmad,[1,a] N.M. Arifin,[1,b] R. Nazar,[2] I. Pop[3]

[1] Institute for Mathematical Research, Universiti Putra Malaysia
43400 UPM Serdang, Selangor, Malaysia,
[a] syakilaahmad@yahoo.com, [b] norihan@math.upm.edu.my

[2] School of Mathematical Sciences, Universiti Kebangsaan Malaysia
43600 UKM Bangi, Selangor, Malaysia, rmn72my@yahoo.com

[3] Faculty of Mathematics, University of Cluj,
CP 253, R-3400 Cluj, Romania, pop.ioan@yahoo.co.uk

Abstract. The problem of steady laminar forced convection boundary layer flow of an incompressible viscous fluid over a moving thin needle with variable heat flux is considered. The governing boundary layer equations are first transformed into non-dimensional forms. These equations are then transformed into similarity equations using the similarity variables, which are solved numerically using an implicit finite-difference scheme known as the Keller-box method. The solutions are obtained for a blunt-nosed needle (m=0). Numerical computations are carried out for various values of the dimensionless parameters of the problem, which include the Prandtl number Pr and the parameter a representing the needle size. It has been found that the wall temperature is significantly influenced by both parameter a and Prandtl number Pr . However, the Prandtl number has no effect on the flow characteristics due to the decoupled boundary layer equations.

Keywords. Boundary layer flow, Moving thin needle, Variable heat flux

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_4,

4.1 Introduction

Thin needle is a body of revolution whose diameter is of the same order as the velocity or thermal boundary layers that it develops. By appropriately varying the radius of the needle, the partial differential boundary layer equations admit similarity solutions, which are more revealing than the direct numerical integration of the partial differential equations. As it is well known, an important aspect of experimental studies for the flow and heat transfer characteristics is the measurements of velocity and temperature profiles of the flow field. The probe of the measuring devices, such as a hot wire anemometer or shielded thermocouple, is often a very thin wire or needle. Meanwhile, boundary layer behavior over moving solid surface is an important type of flow occurring in a number of engineering processes. Aerodynamic extrusion of plastic sheet, cooling of an infinite metallic plate in a cooling bath, the boundary layer along a liquid film in condensation processes, and a polymer sheet or filament extruded continuously from a dye, or a long thread traveling between a feed roll and a windup roll, are examples of practical applications of continuous surfaces (see [1, 16]). From an industrial point of view, the wall shear stress distribution is perhaps the most important parameter in this type of flow because it directly determines the driving force (or torque) required to withdraw the surface (see [15]). Therefore, the detailed analysis of the flow over such moving slender needle-shaped bodies is of considerable practical interest.

The problems of forced, free, and mixed convection boundary layer flows over thin needles have been investigated by many researchers. Chen and Smith [6], Narain and Uberoi [13,14], Chen [5], Lee et al. [12], and Ahmad et al. [4] have studied various aspects of this problem. Wang [17] has studied the problem of mixed convection boundary layer flow on a vertical adiabatic thin needle with a concentrated heat source at the tip of the needle. This situation may be applied, for example, to a stick burning at the bottom end. Agarwal et al. [2] have investigated numerically the momentum and thermal boundary layers for power-law fluids over a thin needle under wide ranges of kinematic and physical conditions. We also notice to this end that Gorla [8–10] has studied the boundary layer flow in the vicinity of an axisymmetric stagnation point on a circular cylinder placed in a Newtonian or in a micropolar fluid.

All studies mentioned above on forced, free, or mixed convection boundary layer flows over thin needles refer to fixed needles immersed in a viscous and incompressible fluid. However, the solutions for mixed convection boundary layer flow past a vertical moving thin needle in a quiescent fluid with variable heat flux have been reported recently by Ahmad et al. [3]. The aim of this chapter is to study the problem of steady forced convection boundary layer flow over a moving thin needle with variable heat flux in a quiescent fluid. It should also be mentioned that due to entrainment of the ambient fluid, this flow situation represents an intrinsically different class of boundary layer flows, which have substantially different type of solutions as compared to the case of a static needle. By the similarity transformation, the partial differential equations governing the flow and temperature fields are reduced to ordinary differential equations, which are solved numerically using an implicit finite-difference scheme called the Keller-box method. The influences of the needle size and the Prandtl number on the flow and heat transfer characteristics are presented in graphical form.

4.2 Mathematical formulation

Consider a steady laminar boundary layer flow of an incompressible viscous fluid over a moving thin needle in a bulk fluid at a constant temperature T_∞. Figure 4.1 shows the slender paraboloid needle whose radius is described by $\bar{r} = \bar{R}(\bar{x})$, where $\bar{x}$ and $\bar{r}$ are the axial and radial coordinates, respectively, with the $\bar{x}$- axis measured from the needle leading edge.

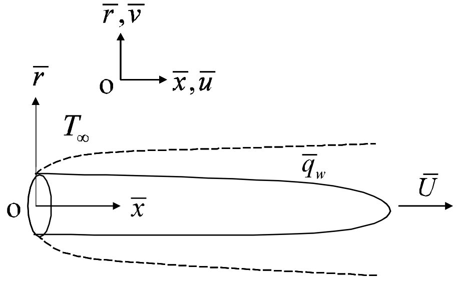


Fig. 4.1. Physical model and coordinate system

The needle is considered to be thin when its thickness does not exceed that of the boundary layer over it. Under this assumption, the effect of transverse curvature is of importance, but the pressure variation along the surface due to the presence of the needle can be neglected (see [11]). It is assumed that the needle moves horizontally with the velocity $\bar{U}(\bar{x})$ and is subjected to a variable surface heat flux $\bar{q}_w(\bar{x})$. Under the boundary layer approximations, the basic boundary layer equations written in cylindrical coordinates are

$$\frac{\partial}{\partial \bar{x}}(\bar{r}\bar{u})+\frac{\partial}{\partial \bar{r}}(\bar{r}\bar{v})=0 \tag{4.1}$$

$$\bar{u}\frac{\partial \bar{u}}{\partial \bar{x}}+\bar{v}\frac{\partial \bar{u}}{\partial \bar{r}}=\frac{\upsilon}{\bar{r}}\frac{\partial}{\partial \bar{r}}\left(\bar{r}\frac{\partial \bar{u}}{\partial \bar{r}}\right) \tag{4.2}$$

$$\bar{u}\frac{\partial \bar{T}}{\partial \bar{x}}+\bar{v}\frac{\partial \bar{T}}{\partial \bar{r}}=\frac{\alpha}{\bar{r}}\frac{\partial}{\partial \bar{r}}\left(\bar{r}\frac{\partial \bar{T}}{\partial \bar{r}}\right) \tag{4.3}$$

where $\bar{u}$ and $\bar{v}$ are the velocity components along the $\bar{x}$- and $\bar{r}$- axes, respectively, $\bar{T}$ is the local fluid temperature, υ is the kinematic viscosity, and α is the constant thermal diffusivity of the fluid. We assume that the boundary conditions of Eqs. (4.1), (4.2) and (4.3) are

$$\begin{aligned}&\bar{v}=0, \bar{u}=\bar{U}(\bar{x}), \frac{\partial \bar{T}}{\partial \bar{r}}=-\frac{\bar{q}_w(\bar{x})}{k} \quad \text{at} \quad \bar{r}=\bar{R}(\bar{x})\\ &\bar{u}\rightarrow 0, \bar{T}\rightarrow T_\infty \quad \text{as} \quad \bar{r}\rightarrow\infty,\end{aligned} \tag{4.4}$$

where $\bar{R}(\bar{x})$ prescribes the surface shape of the axisymmetric body.

We introduce now the following non-dimensional variables:

$$\begin{aligned}&x=\frac{\bar{x}}{L}, r=Re^{1/2}\left(\frac{\bar{r}}{L}\right), u=\frac{\bar{u}}{U_0}, v=Re^{1/2}\left(\frac{\bar{v}}{U_0}\right), U(x)=\frac{\bar{U}(\bar{x})}{U_0}\\ &R(x)=Re^{1/2}\left(\frac{\bar{R}(\bar{x})}{L}\right), q_w(x)=\frac{\bar{q}_w(\bar{x})}{q_0}, T=\frac{kRe^{1/2}\left(\bar{T}-T_\infty\right)}{q_0 L}\end{aligned} \tag{4.5}$$

where L is a characteristic length of the needle, U_0 is the characteristic velocity, q_0 is the characteristic heat flux, and Re is the Reynolds number given by

$$Re=\frac{U_0 L}{\upsilon} \tag{4.6}$$

Substituting variables Eq. (4.5) into Eqs. (4.1), (4.2) and (4.3), we get

$$\frac{\partial}{\partial x}(ru)+\frac{\partial}{\partial r}(rv)=0 \tag{4.7}$$

$$u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial r}=\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial u}{\partial r}\right) \tag{4.8}$$

$$u\frac{\partial T}{\partial x}+v\frac{\partial T}{\partial r}=\frac{1}{Pr}\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial T}{\partial r}\right) \tag{4.9}$$

where Pr is the Prandtl number, with the boundary conditions Eq. (4.4) becoming

$$\begin{aligned} &v=0, u=U(x), \frac{\partial T}{\partial r}=-q_w(x) \quad \text{at} \quad r=R(x) \\ &u=0, T=0 \quad \text{as} \quad r\rightarrow\infty \end{aligned} \tag{4.10}$$

In order that Eqs. (4.7), (4.8) and (4.9) become similar, we take

$$U(x)=x^m, \quad q_w(x)=x^{(5m-1)/2} \tag{4.11}$$

where m is a constant. We introduce now the following similarity variables:

$$\psi=xf(\eta), \quad T(x)=x^{2m}\theta(\eta) \tag{4.12}$$

where

$$\eta=x^{m-1}r^2 \tag{4.13}$$

and ψ is the stream function which is defined in the usual way as $u=(1/r)\partial\psi/\partial r$ and $v=-(1/r)\partial\psi/\partial x$. The introduction of the stream function automatically satisfies the continuity equation (4.7). The surfaces of constant $\eta=a$, where a is a non-dimensional constant and is numerically small for a slender body, correspond to the surfaces of revolution. Setting $\eta=a$, Eq. (4.13) prescribes both shape and size of the body with its surface given by

$$R(x)=a^{1/2}x^{(1-m)/2} \tag{4.14}$$

Of practical interest are pointed bodies and cylinders for which we must have, from Eq. (4.14), $m\le 1$. For example, the body is a cylinder when $m=1$, a paraboloid when $m=0$, and a cone when $m=-1$. Substituting Eqs. (4.12) and (4.13) into Eqs. (4.8) and (4.9), we get the following ordinary differential equations:

$$2(\eta f'')'+ff''-mf'^2=0 \tag{4.15}$$

$$\frac{2}{Pr}(\eta\theta')' + f\theta' - 2mf'\theta = 0 \tag{4.16}$$

subject to the boundary conditions

$$f(a) = \frac{(1-m)a}{2}, \quad f'(a) = \frac{1}{2}, \quad f'(\infty) = 0 \tag{4.17}$$
$$\theta'(a) = -\frac{1}{2}a^{-1/2}, \quad \theta(\infty) = 0$$

where primes denote differentiation with respect to η.

The physical quantities of interest are the skin friction coefficient C_f and the wall temperature T_w which are defined as

$$C_f = \frac{\tau_w}{\rho\bar{U}^2/2}, \quad T_w = \frac{kRe^{1/2}\left(\bar{T}_w - T_\infty\right)}{q_0 L} \tag{4.18}$$

where $\bar{T}_w$ is the dimensional wall temperature and the skin friction τ_w is given by

$$\tau_w = \mu\left(\frac{\partial\bar{u}}{\partial\bar{r}}\right)_{\bar{r}=\bar{R}(\bar{x})} \tag{4.19}$$

Using Eqs. (4.5), (4.12), (4.13), and Eq. (4.19), we get

$$C_f Re_x^{1/2} = 8a^{1/2}f''(a), \quad T_w = x^{2m}\theta(a) \tag{4.20}$$

where Re_x is the local Reynolds number which is defined as

$$Re_x = \frac{\bar{U}(\bar{x})\bar{x}}{\upsilon} \tag{4.21}$$

4.3 Results and discussion

Generally, as mentioned in the previous section, the body is a cylinder when $m = 1$, a paraboloid when $m = 0$, and a cone when $m = -1$ (see Chen [5]). However, when a solid object of any shape, such as a needle in this present problem, exhibits a rectilinear translational motion through a fluid medium, then all parts of the solid object (needle) must have the same velocity. This implies that the velocity of its surface U has to be a constant and therefore only the value $m = 0$ should be considered in this chapter. However, the only exception one may think of is an elastic body. The most typical example is an elastic sheet that is being stretched, in which the sheet velocity varies along the sheet. Therefore, in the present problem, the results for $m = -1$ and $m = +1$ are not of any physical relevance since Eq. (4.11) is inconsistent with a solid needle. Still, these results are solutions of the mathematical problem posed, but without any physical realism. According to Eq. (4.14) the value of $m = 0$ corresponds to a blunt-nosed needle or a paraboloid with $R(x) = a^{1/2}x^{1/2}$.

The system of decoupled ordinary differential equations (4.15) and (4.16) subject to the boundary conditions Eq. (4.17) has been solved numerically using an implicit finite-difference method known as the Keller-box scheme as described in the book by Cebeci and Bradshaw [7] for $m = 0$ (a blunt-nosed needle with variable heat flux) and some values of the governing parameter a (in the range of $0.001 \leq a \leq 0.1$). The solution is obtained in the following four steps:

1. reduce Eqs. (4.15) and (4.16) to a first-order system,
2. write the difference equations using central differences,
3. linearize the resulting algebraic equations by Newton's method and write them in matrix-vector form,
4. solve the linear system by the block-tridiagonal-elimination technique.

We consider that the needle moves in a fluid with different Prandtl numbers, i.e., Pr varies in the range of $0.01 \leq Pr \leq 100$. It is worth mentioning that small values of Pr ($\ll 1$) physically correspond to liquid metals, which have high thermal conductivity but low viscosity, while $Pr \sim 1$ corresponds to diatomic gases including air. On the other hand, large values of Pr ($\gg 1$) correspond to high-viscosity oils and $Pr = 6.8$ corresponds to water at room temperature. Results are presented in six figures.

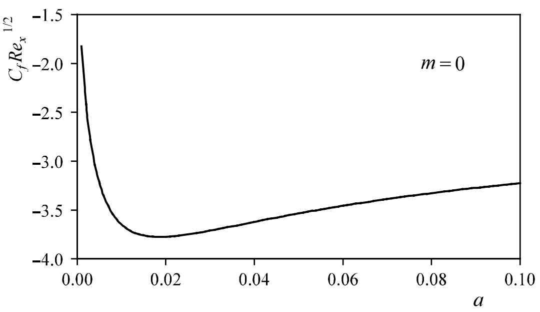


Fig. 4.2. Variation of the skin friction coefficient with a for various Pr when $m = 0$

The variations with a of the skin friction coefficient $C_f Re_x^{1/2}$ and the wall temperature T_w, given by expressions Eq. (4.20) for $m = 0$, are shown in Figs. 4.2 and 4.3 for $Pr = 0.01, 0.1, 0.7, 1, 6.8,$ and 10. Due to the decoupled boundary layer equations (4.15) and (4.16), it is seen from Fig. 4.2 that there is only a unique skin friction coefficient for all considered value of Pr at different values of a. It can also be seen from Fig. 4.2 that the skin friction coefficient decreases with the increase of a for $0.001 \le a \le a_{\text{min}}$ where a_{min} is the value of a when the skin friction coefficient is minimum. Figure 4.2 also shows that for $a_{\text{min}} < a \le 0.1$, the skin friction coefficient increases with the increase of a. It is worth mentioning that the negative sign of the skin friction coefficient in Fig. 4.2 physically implies that the fluid produces a dragging force on the surface.

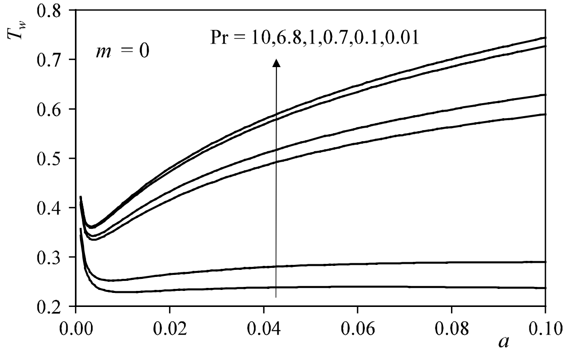


Fig. 4.3. Variation of the wall temperature with a for various Pr when $m = 0$

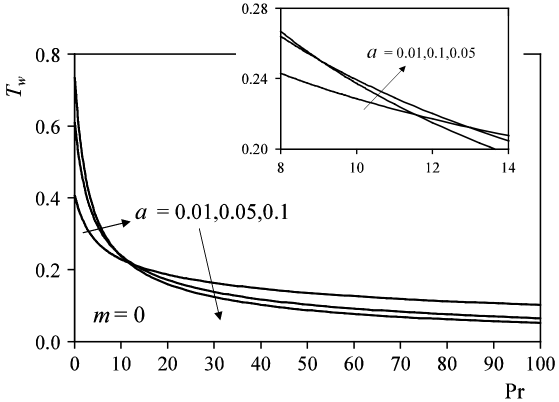


Fig. 4.4. Variation of the wall temperature with Pr for various values of a when $m = 0$

Figures 4.3 and 4.4 show the variation with a and *P*r, respectively, for the wall temperature of a blunt-nosed needle with variable heat flux ($m = 0$). It can be seen from those figures that at any fixed values of a, the wall temperature decreases as Pr increases. Physically, this is because the thermal diffusivity in boundary layer becomes higher as Pr increases. Figure 4.3 shows that for a fixed value of Pr, the wall temperature decreases with the increase of a for $0.001 \le a \le a_{\min}(Pr)$ where $a_{\min}(Pr)$ is the value of a when the wall temperature is minimum and depending on Pr. On the other hand, it is seen from Fig. 4.3 that for $a_{\min}(Pr) < a \le 0.1$, the wall temperature increases when a increases. Furthermore, it can be seen from Fig. 4.4 that for $Pr < Pr_{in}^{0.05,0.1}$ where $Pr_{in}^{0.05,0.1}$ is the value of Pr when the curves $a = 0.05$ and $a = 0.1$ intersect, the wall temperature increases with the increase of a. On the other hand, Fig. 4.4 also shows that the wall temperature decreases with the increase of a when $Pr > Pr_{in}^{0.01,0.05}$ where $Pr_{in}^{0.01,0.05}$ is the value of Pr when the curves $a = 0.01$ and $a = 0.05$ intersect.

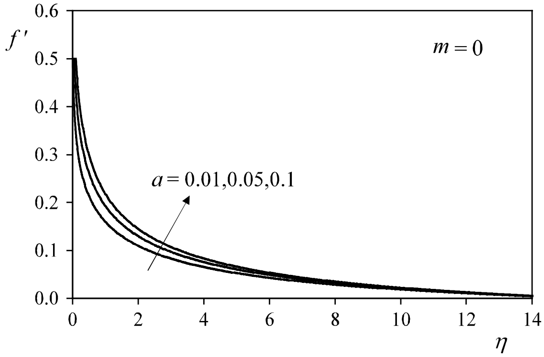


Fig. 4.5. Velocity profiles for various values of a with $m = 0$

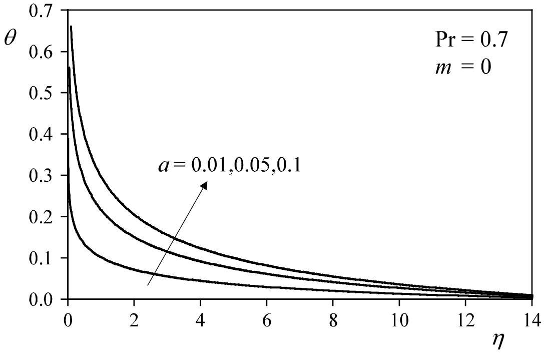


Fig. 4.6. Temperature profiles for various values of a with $Pr = 0.7$ and $m = 0$

The axial velocity profiles $f'(\eta)$ and the non-dimensional temperature profiles $\theta(\eta)$ for a blunt-nosed needle with variable heat flux ($m=0$) are plotted versus η in Figs. 4.5 and 4.6, respectively, for three needle sizes, namely $a=0.01, 0.05,$ and 0.1. Figure 4.5 shows that at any fixed value of a, there is only a unique velocity profile for all values of *Pr*. It is seen from Figs. 4.5 and 4.6 that the velocity and thermal boundary layer thicknesses increase with the increase of the needle size a. An inspection of these figures clearly shows that the thinner the needle, the smaller the value of η for the free stream conditions to be attained, i.e., the boundary layer thickness decreases with the decreasing values of a. It can also be clearly seen from Fig. 4.5 that $f''(\eta) \to 0$ as $\eta \to \infty$, i.e., the shear stress vanishes outside the momentum boundary layer.

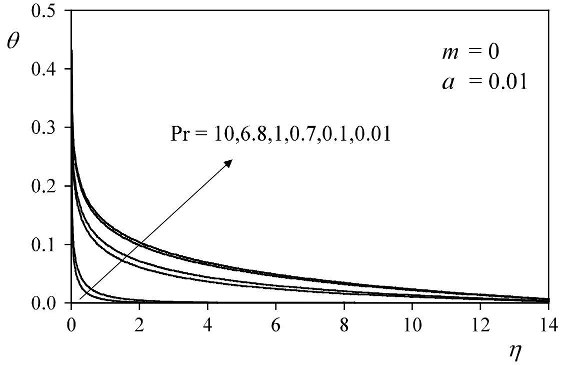


Fig. 4.7. Temperature profiles for various *Pr* with $m=0$ and $a=0.01$

Figure 4.7 displays the non-dimensional temperature profiles $\theta(\eta)$ for a blunt-nosed needle with variable heat flux ($m=0$) for various values of *Pr* ($Pr=0.01, 0.1, 0.7, 1, 6.8, 10$) and $a=0.01$. It is shown that the temperature profile decreases as *Pr* increases. It is also shown that the thermal boundary layer thickness decreases with an increase in *Pr*. Physically, this is because, as *Pr* increases, the thermal diffusivity decreases. This leads to the decrease of the energy transfer ability that reduces the thermal boundary layer.

4.4 Conclusions

The problem of steady laminar forced convection boundary layer flows of an incompressible viscous fluid over a moving thin needle in an ambient fluid is studied. Calculations are carried out for a blunt-nosed needle with variable heat flux ($m = 0$), which moves in a fluid with a wide range of the Prandtl numbers ($0.01 \leq Pr \leq 100$). The numerical results are also obtained for various values of the dimensionless parameters, which include the Prandtl number Pr and the parameter a representing the needle size. The results show that the shape and the size of the needles have strong effects on the velocity and thermal characteristics of the problem. Generally, it can be concluded that the wall temperature and the temperature profiles are significantly influenced by the considered parameters. However, the Prandtl number has no effect on the local skin friction coefficient and the velocity profiles due to the decoupled boundary layer equations.

References

1. Abraham JP, Sparrow EM (2005) Friction drag resulting from the simultaneous imposed motions of a freestream and its bounding surface. Int J Heat Fluid Flow 26:289–295
2. Agarwal M, Chhabra RP, Eswaran V (2002) Laminar momentum and thermal boundary layers of power-law fluids over a slender cylinder. Chem Engng Sci 57:1331–1341
3. Ahmad S, Arifin NM, Nazar R, Pop I (2008a) Mixed convection boundary layer flow along vertical moving thin needles with variable heat flux. Heat Mass Transf 44:473–479
4. Ahmad S, Arifin NM, Nazar R, Pop I (2008b) Mixed convection boundary layer flow along vertical thin needles: assisting and opposing flows. Int Comm Heat Mass Transf 35:157–162
5. Chen JLS (1987) Mixed convection flow about slender bodies of revolution. J Heat Transf 109:1033–1036
6. Chen JLS, Smith TN (1978) Forced convection heat transfer from nonisothermal thin needles. J Heat Transf 100:358–362
7. Cebeci T, Bradshaw P (1988) Physical and computational aspects of convective heat transfer. Springer, New York
8. Gorla RSR (1979) Unsteady viscous flow in the vicinity of an axisymmetric stagnation point on a circular cylinder. Int J Engng Sci 17:87–93
9. Gorla RSR (1990) Boundary layer flow of a micropolar fluid in the vicinity of an axisymmetric stagnation point on a cylinder. Int J Engng Sci 28:145–152
10. Gorla RSR (1993) Mixed convection in an axisymmetric stagnation flow on a vertical cylinder. Acta Mech 99:113–123
11. Lee LL (1967) Boundary layer over a thin needle. Phys Fluids 10:820–822
12. Lee SL, Chen TS, Armaly BF (1987) Mixed convection along vertical cylinders and needles with uniform surface heat flux. J Heat Transf 109:711–716

13. Narain JP, Uberoi MS (1972) Combined forced and free-convection heat transfer from vertical thin needles in a uniform stream. Phys Fluids 15:1879–1882
14. Narain JP, Uberoi MS (1973) Combined forced and free-convection over thin needles. Int J Heat Mass Transf 16:1505–1511
15. Sadeghy K, Sharifi M (2004) Local similarity solution for the flow of a "second-grade" viscoelastic fluid above a moving plate. Int J Non-Linear Mech 39:1265–1273
16. Sparrow EM, Abraham JP (2005) Universal solutions for the streamwise variation of the temperature of a moving sheet in the presence of a moving fluid. Int J Heat Mass Transf 48:3047–3056
17. Wang CY (1990) Mixed convection on a vertical needle with heated tip. Phys Fluids A 2:622–625

Chapter 5

The parallel three-processor fifth-order diagonally implicit Runge–Kutta methods for solving ordinary differential equations

U.K.S. Din[1], F. Ismail[2], M. Suleiman[2], Z.A. Majid[2], M. Othman[3]

[1] School of Mathematical Sciences, Faculty of Science & Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia, ummul@ukm.my
[2] Department of Mathematics, Faculty of Science,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia, fudziah_i@yahoo.com.my, mohameds@science.upm.edu.my, zanamajid@hotmail.com
[3] Department of Communication Technology & Network,
Faculty of Computer Science & Information Technology,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia,
mothman@fsktm.upm.edu.my

Abstract. Fifth-order diagonally implicit Runge–Kutta methods with a modified sparsity structure suitable for parallel implementations on three processors are developed. The efficiency of the methods in terms of accuracy to solve a standard set of problems is compared to an established method. From the results we can conclude the new methods are comparable to the existing method.

Keywords. Fifth-order diagonally implicit Runge–Kutta methods, Sparsity structure, Parallelism, Ordinary differential equations

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_5,

5.1 Introduction

The primary objective in applying parallelism in numerical computation is the significant reduction in time which appears to be one of the contributing costs in computer execution. Iserles and Nørsett [7] and Jackson and Nørsett [9] have presented the idea of developing parallel Runge–Kutta methods through sparsity structure which is considered as one of the best parallel designs for Runge–Kutta methods. The structure allows the evaluation of the functions with independent arguments to be computed on different processors at the same time as one single evaluation on one processor. Furthermore the structures are meant for implicit (IRK) and diagonally implicit Runge–Kutta (DIRK) methods where a higher accuracy could be achieved as well as it can be applied to solve stiff ordinary differential equations (ODEs).

Previous methods using the sparsity structure are fourth-order methods suitable to be implemented on two processors [10,8,14,9,3]. In this section, we present fifth-order Runge–Kutta methods, where a pattern of DIRK that was suggested by Iserles and Nørsett [7] and Jackson and Nørsett [9] is developed to permit the implementation of parallel computation using three processors.

5.2 The Runge–Kutta method

The initial value problem (IVP) for a system of s first-order ODEs is defined by

$$y'(x) = f(x, y), \qquad x \in [a, b], \qquad y(a) = y_0 \tag{5.1}$$

The general s-stage Runge–Kutta method for problem (5.1) is defined by

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i, \tag{5.2}$$

where

$$k_i = f\left(x_n + c_i h, y_n + h \sum_{j=1}^{s} a_{ij} k_j \right), \qquad i = 1, 2, \ldots, s \tag{5.3}$$

assuming the following holds

$$c_i = \sum_{j=1}^{s} a_{ij} \ , \qquad\qquad i = 1, 2, \ldots, s \tag{5.4}$$

In *Butcher' array* [4] the coefficients in Equations (5.3) and (5.4) are written as

$$\begin{array}{c|ccccc}
c_1 & a_{11} & a_{12} & a_{13} & \cdots & a_{1s} \\
c_2 & a_{21} & a_{22} & a_{23} & \cdots & a_{2s} \\
\vdots & & \vdots & & & \vdots \\
c_s & a_{s1} & a_{s2} & a_{s3} & \cdots & a_{ss} \\
\hline
 & b_1 & b_2 & b_3 & \cdots & b_s
\end{array}$$

or simply as

$$\begin{array}{c|c}
c & A \\
\hline
 & b^T
\end{array}$$

with the s-dimensional vectors c and b and the $s \times s$ matrix A denoted by $c = [c_1, c_2, c_3, \ldots, c_s]^T$, $b = [b_1, b_2, b_3, \ldots, b_s]^T$, and $A = \left[a_{ij} \right]$, respectively. The method is said to be explicit if $a_{ij} = 0$ for $i \leq j$, semi-implicit if $a_{ij} = 0$ for $i < j$, and fully implicit otherwise.

5.2.1 The workload and the needs for parallel computing

Shampine [12] has noted that the cost of an integration is conventionally measured by the number of function evaluations required and in Runge–Kutta methods this is done at every stage which is represented by k_i in (5.3). This means that the cost would increase if a higher number of k is involved and would continue to increase if we deal with a large problem. An example of problem that requires a huge number of calculations is predicting the motion of the astronomical bodies in space, better known as the N-body problem. The N-body problem also appears in modeling chemical and biological systems at the molecular level and takes enormous computational power. Basically there are two reasons for using parallel computing: to save time and to solve large problem. By having a few processors, or sometimes referred to as *workers* or *laborers* or *slaves*, to do calculations concurrently, large problems could be solved in shorter time.

5.2.2 Parallelism of the method

The theory expressed in Iserles and Nørsett [7] depicted through directed digraphs explicitly showed groups of stages which are independent of each other. The sparsity pattern as shown in Fig. 5.1 together with its digraph has been chosen in developing our method as we intend to implement the method on three processors.

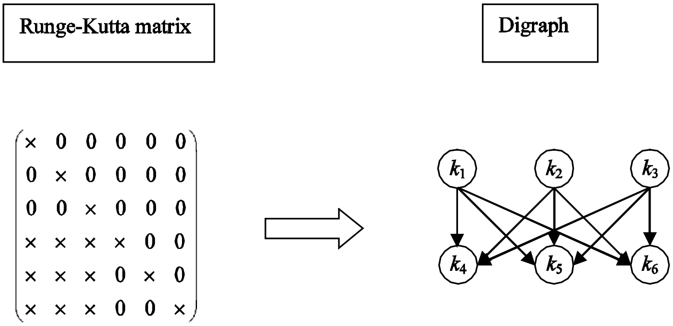


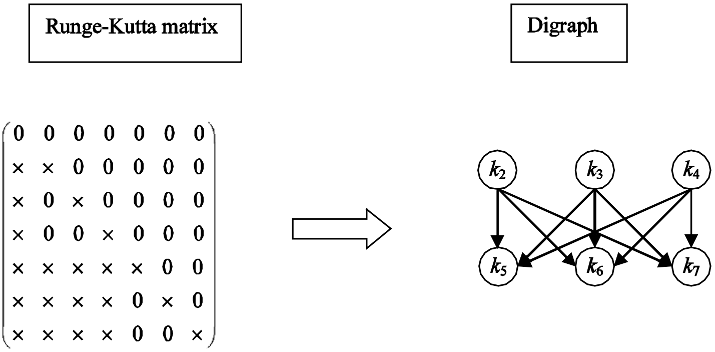
Fig. 5.1. Sparsity structure and digraph for Runge–Kutta methods with six stages for three processors

Initial efforts in deriving the method are based wholly on the array in Fig. 5.2 which represents the Runge–Kutta method for the sparsity structure in Fig. 5.1. However, we found that with certain a_{ij} assigned to zero, the freedom in manipulating the system of equations obtained from the order conditions is limited, leading to the lack of unknowns compared to the number of equations. These would make it impossible to find the solutions.

$$
\begin{array}{c|cccccc}
c_1 & \alpha & & & & & \\
c_2 & 0 & \beta & & & & \\
c_3 & 0 & 0 & \gamma & & & \\
c_4 & a_{51} & a_{52} & a_{53} & \alpha & & \\
c_5 & a_{61} & a_{62} & a_{63} & 0 & \beta & \\
c_6 & a_{71} & a_{72} & a_{73} & 0 & 0 & \gamma \\
\hline
 & b_1 & b_2 & b_3 & b_4 & b_5 & b_6
\end{array}
$$

Fig. 5.2. Butcher's array for Runge–Kutta methods with six stages for three processors

To overcome this situation, we decided to add an explicit first stage to the method without changing the original sparsity structure. The new matrix is shown in Fig. 5.3 together with its digraph.



$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \times & \times & 0 & 0 & 0 & 0 & 0 \\ \times & 0 & \times & 0 & 0 & 0 & 0 \\ \times & 0 & 0 & \times & 0 & 0 & 0 \\ \times & \times & \times & \times & \times & 0 & 0 \\ \times & \times & \times & \times & 0 & \times & 0 \\ \times & \times & \times & \times & 0 & 0 & \times \end{pmatrix}$$

Fig. 5.3. The modified sparsity structure and digraph for Runge–Kutta methods for three processors

5.3 Derivation of fifth-order parallel Runge–Kutta methods for three processors

In order to derive a fifth-order diagonally implicit Runge–Kutta method, 17 equations have to be satisfied. The equations associated with the order of the method are given in Table 5.1.

We used four assumptions [4,5] to reduce and simplify the equations so that it would be easier to solve. The assumptions are

$$\sum_{i=1}^{7} b_i c_i^{\,k} = \frac{1}{k+1} \qquad \text{for } k\text{=0,1,2,3,4} \tag{5.5}$$

$$\sum_{j=1}^{7} a_{ij} c_j = \frac{1}{2} c_i^{\,2} \qquad \text{for } i\text{=1,2,…,7} \tag{5.6}$$

$$\sum_{j=1}^{7} a_{ij} c_j^{\,2} = \frac{1}{3} c_i^{\,3} \qquad \text{for } i\text{=1,2,…,7} \tag{5.7}$$

$$\sum_{i=1}^{7} b_i a_{ij} = b_j (1 - c_j) \qquad \text{for } j\text{=1,2,…,7} \tag{5.8}$$

According to Butcher [5], assumption (5.5) usually denoted by $B(\eta)$ is necessary to be satisfied for a method to be of order η. Therefore

Equations (5.9), (5.10), (5.11), (5.13), and (5.17) have to be satisfied. Equation (5.9) is also known as the consistency condition.

Using Equations (5.6) and (5.7), we removed Equations (5.12), (5.14), (5.15), (5.16), (5.18), (5.19), (5.20), (5.22), (5.23), (5.24), and (5.25). Generally, Equations (5.6) and (5.7) are meant for *i* = 1,2,…,*s*. In fact some unknowns could be determined when further inspection is done. For example when *i* = 2, Equation (5.6) is simplified to

$$\alpha c_2 = \frac{c_2^{\ 2}}{2}$$

$$\Rightarrow c_2 = 0 \text{ or } c_2 = 2\alpha$$

But further inspection of Equation (5.7) for *i* = 2 shows that c_2 is equal to either zero or 3α. Since we do not want c_2 to be zero, we assigned b_2 to zero. The same thing arises for *i*=3 and 4; therefore, we have b_3 and b_4 also equal to zero.

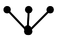
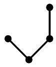
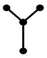
With the same arguments we imposed

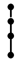
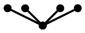
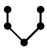
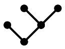
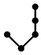
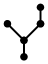
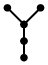
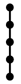
$$\sum b_i a_{i2} = 0, \qquad \sum b_i c_i a_{i2} = 0$$

$$\sum b_i a_{i3} = 0, \qquad \sum b_i c_i a_{i3} = 0$$

$$\sum b_i a_{i4} = 0, \qquad \sum b_i c_i a_{i4} = 0$$

Table 5.1. Equations of order conditions for Runge–Kutta methods of order 5

Order of the tree	Tree	Elementary weights	
1		$\sum b_i = 1$	(5.9)
2		$\sum b_i c_i = \frac{1}{2}$	(5.10)
3		$\sum b_i c_i^2 = \frac{1}{3}$	(5.11)
3		$\sum b_i a_{ij} c_j = \frac{1}{6}$	(5.12)
4		$\sum b_i c_i^3 = \frac{1}{4}$	(5.13)
4		$\sum b_i c_i a_{ij} c_j = \frac{1}{8}$	(5.14)
4		$\sum b_i a_{ij} c_j^2 = \frac{1}{12}$	(5.15)

4		$\sum b_i a_{ij} a_{jk} c_k = \frac{1}{24}$	(5.16)
5		$\sum b_i c_i^4 = \frac{1}{5}$	(5.17)
5		$\sum b_i c_i^2 a_{ij} c_j = \frac{1}{10}$	(5.18)
5		$\sum b_i a_{ij} c_j a_{ik} c_k = \frac{1}{20}$	(5.19)
5		$\sum b_i c_i a_{ij} c_j^2 = \frac{1}{15}$	(5.20)
5		$\sum b_i a_{ij} c_j^3 = \frac{1}{20}$	(5.21)
5		$\sum b_i c_i a_{ij} a_{jk} c_k = \frac{1}{30}$	(5.22)
5		$\sum b_i a_{ij} c_j a_{jk} c_k = \frac{1}{40}$	(5.23)
5		$\sum b_i a_{ij} a_{jk} c_k^2 = \frac{1}{60}$	(5.24)
5		$\sum b_i a_{ij} a_{jk} a_{kl} c_l = \frac{1}{120}$	(5.25)

As for Equation (5.21), the last assumption (5.8) will be applied where it will give values for c_5, c_6, and c_7 which are $1-\alpha, 1-\beta$, and $1-\gamma$, respectively. The simplifying processes leave us with a total of 17 equations and 18 unknowns to be solved. We solved the equations using Mathematica and have come out with two sets of solution. The first solution is obtained when we set the value of α=0.25 and β=0.5 which will be denoted as P3DIRK5(i) and for the second solution α is 0.5 and β is 0.75, denoted as P3DIRK5(ii). Both solutions are given in Tables 5.2 and 5.3, respectively.

Table 5.2. The solution for P3DIRK5(i)

0	**0**						
0.33333	**0.08333**	**0.25**					
0.4	**−0.1**	**0**	**0.5**				
0	**−0.4**	**0**	**0**	**0.4**			
0.75	**−0.796878**	**1.68743**	**−1.1718**	**0.78125**	**0.25**		
0.5	**−0.631947**	**1.49994**	**−1.56244**	**0.694444**	**0**	**0.5**	
0.6	**−1.01**	**2.15991**	**−1.94991**	**1**	**0**	**0**	**0.4**
	0.12963	**0**	**0**	**0**	**1.18519**	**2**	**−2.31481**

Table 5.3. The solution for P3DIRK5(ii)

0	**0**						
0.5	**0**	**0.5**					
0.4	**–0.35**	**0**	**0.75**				
0	**–0.13333**	**0**	**0**	**0.13333**			
0.5	**2.74199**	**–0.666667**	**0.520833**	**–2.59615**	**0.5**		
0.25	**–1.6278**	**0.416667**	**–0.911458**	**1.6226**	**0**	**0.75**	
0.866667	**–0.852469**	**0.25679**	**0.329012**	**1**	**0**	**0**	**0.13333**
	0.0897436	**0**	**0**	**0**	**0.30303**	**0.288288**	**0.318938**

5.4 Numerical experiments

All problems tested are non-stiff problems. We compare the accuracy of the methods to fifth-order DIRK methods by Al-Rabeh [1].

Problem1:

$y' = -y$

$y(0) = 1$, $0 \le x \le 20$.

Exact solution: $y(x) = e^{-x}$.

Source: Artificial problem.

Problem 2:

$y_1' = -y_1 - \sqrt{3} y_2$

$y_2' = \sqrt{3} y_1 - y_2$

$y_1(0) = 1$, $y_2(0) = 0$, $0 \le x \le 20$.

Exact solutions: $y_1(x) = e^{-x} \cos \sqrt{3} x$, $y_2(x) = e^{-x} \sin \sqrt{3} x$.

Source: Tam [13].

Problem 3:

$y_1' = y_2 - y_1$

$y_2' = y_1 - 2y_2 + y_3$

$y_3' = y_2 - y_3$

$y_1(0) = 2$, $y_2(0) = 0$, $y_3(0) = 1$, $0 \le x \le 20$.

Exact solutions:

$$y_1(x) = \frac{1}{2}e^{-3x} + 1 + \frac{1}{2}e^{-x}$$

$$y_2(x) = 1 - e^{-3x}$$

$$y_3(x) = 1 + \frac{1}{2}e^{-3x} - \frac{1}{2}e^{-x} \quad .$$

Source: Shampine [11].

Problem 4:

$$y_1' = y_2$$

$$y_2' = -y_3$$

$$y_3' = y_4$$

$$y_4' = y_2 + 2e^x$$

$y_1(0) = 0$, $y_2(0) = -2$, $y_3(0) = 0$, $y_4(0) = 2$, $0 \le x \le 10$.

Exact solutions:

$y_1(x) = -e^x + e^{-x}$, $y_2(x) = -e^x - e^{-x}$,

$y_3(x) = e^x - e^{-x}$, $y_4(x) = e^x + e^{-x}$.

Source: Bronson [2].

Problem 5:

$y_1' = y_3$, $y_2' = y_4$, $y_3' = -\frac{y_1}{r^3}$, $y_4' = -\frac{y_2}{r^3}$,

$r = \sqrt{y_1^2 + y_2^2}$,

$y_1(0) = 1$, $y_2(0) = 0$, $y_3(0) = 0$, $y_4(0) = 1$, $0 \le x \le 75$.

Exact solutions:

$y_1(x) = \cos x$, $y_2(x) = \sin x$, $y_3(x) = -\sin x$, $y_4(x) = \cos x$.

Source: Dormand [6].

The code for the algorithm to run the method is done in C language. Tables 5.4 and 5.5 show the performance comparison between P3DIRK5(i), P3DIRK5 (ii), and Al-Rabeh in term of maximum error. The step sizes used are 10^{-2}, 10^{-3}, 10^{-4}, and 10^{-5}. The maximum error is defined as

$$\max_{1 \le i \le steps} \left(\left| y_i - y(x_i) \right| \right)$$

where y_i is the computed value and $y(x_i)$ is the true solution of the problems.

Table 5.4. Numerical results for test problems 1–5 using P3DIRK5(i), P3DIRK5(ii), and Al-Rabeh with step sizes 10^{-2} and 10^{-3}

Problem	Method	h=10^{-2}	h=10^{-3}
1	P3DIRK5(i)	1.2898313E-03	1.2956515E-04
	P3DIRK5(ii)	8.9293887E-04	8.9698708E-05
	Al-Rabeh	7.0377347E-04	7.0695342E-05
2	P3DIRK5(i)	1.8299346E-03	1.8475820E-04
	P3DIRK5(ii)	1.2667456E-03	1.2777535E-04
	Al-Rabeh	9.9836453E-04	1.0070508E-04
3	P3DIRK5(i)	1.4575903E-07	1.3666056E-07
	P3DIRK5(ii)	1.3094262E-07	1.2992110E-11
	Al-Rabeh	9.2675797E-10	3.6946779E-11
4	P3DIRK5(i)	2.6018265E-01	3.9015584E-02
	P3DIRK5(ii)	1.7170707E-01	1.7145870E-02
	Al-Rabeh	1.3527754E-01	1.3510061E-02
5	P3DIRK5(i)	5.8036620E-01	5.8702730E-02
	P3DIRK5(ii)	4.0384028E-01	4.0648689E-02
	Al-Rabeh	3.1876621E-01	3.2034277E-02

Table 5.5. Numerical results for test problems 1–5 using P3DIRK5(i), P3DIRK5 (ii), and Al-Rabeh with step sizes 10^{-4} and 10^{-5}

Problem	Method	h=10^{-4}	h=10^{-5}
1	P3DIRK5(i)	1.2962347E-05	1.3004077E-06
	P3DIRK5(ii)	8.9739102E-06	8.9743141E-07
	Al-Rabeh	7.0727168E-06	7.0730351E-07
2	P3DIRK5(i)	1.8667436E-05	2.0585134E-06
	P3DIRK5(ii)	1.2788567E-05	1.2789670E-06
	Al-Rabeh	1.0079157E-05	1.0079559E-06
3	P3DIRK5(i)	1.3629250E-07	1.3625571E-07
	P3DIRK5(ii)	2.0317081E-13	1.7443824E-12
	Al-Rabeh	3.6806225E-11	3.6786352E-11
4	P3DIRK5(i)	2.2239429E-02	2.0562710E-02
	P3DIRK5(ii)	1.7143998E-03	1.7143773E-04
	Al-Rabeh	1.3478727E-03	1.3179881E-04
5	P3DIRK5(i)	5.8451758E-03	5.5989628E-04
	P3DIRK5(ii)	4.0650716E-03	4.0650674E-04
	Al-Rabeh	3.2038271E-03	3.2039211E-04

5.5 Conclusion

From Tables 5.4 and 5.5, it was observed that P3DIRK5(ii) gives better performance in term of accuracy compared to P3DIRK5(i) and as accurate as the method by Al-Rabeh in most of the problems tested except for Problem 3 where as the step size gets smaller P3DIRK5(ii) performed better. As a whole the two new methods are comparable to Al-Rabeh's method. In addition, the methods have the advantage in reducing the cost of computation since it could be implemented in parallel which is very significant when large problems come in hand.

Acknowledgments. The first author acknowledges Universiti Kebangsaan Malaysia and Ministry of Higher Education, Malaysia, for the financial support received throughout this study.

References

1. Al-Rabeh AH (1993) Optimal order diagonally implicit Runge–Kutta methods. BIT 33:620–633
2. Bronson R (1973) Schaum's outline of modern introductory differential equations. McGraw-Hill, New York
3. Burrage K (1995) Parallel and sequential methods for ordinary differential equations. Oxford University Press Inc., New York
4. Butcher JC (1987) The numerical analysis of ordinary differential equations: Runge-Kutta and General Linear methods. John Wiley & Sons, Chichester
5. Butcher JC (2003) Numerical methods for ordinary differential equations. John Wiley & Sons, Chichester
6. Dormand JR (1996) Numerical methods for differential equations: A computational approach. CRC Press, Inc.Boca Raton
7. Iserles A, Nørsett SP (1990) On the theory of parallel Runge-Kutta methods. IMA J. Numer. Anal. 10:463–488
8. Jackson KR (1991) A survey of parallel numerical methods for initial value problems for ordinary differential equations. IEEE Trans. Magn. 27(5):3792–3797
9. Jackson KR, Nørsett SP (1995) The potential for parallelism in Runge-Kutta methods. Part 1: RK formulas in standard form. Siam J. Numer. Anal. 32(1):49–82
10. Nørsett SP, Simonsen HH (1987) Aspects of parallel Runge-Kutta methods. In: Bellen A, Gear CW, Russo E (eds) Numerical methods for ordinary differential equations: Lecture Notes in Mathematics. Springer-Verlag, Berlin, pp 103–107

11. Shampine LF (1980) What everyone solving differential equations numerically should know. In: Gladwell I, Sayers DK (eds) Computational techniques for ordinary differential equations. Academic Press, New York pp 1–17
12. Shampine LF (1994) Numerical solution of ordinary differential equations. Chapman & Hall, New York
13. Tam HW (1992) Two-stage parallel methods for the numerical solution of ordinary differential equations. Siam J. Sci. Stat. Comput. 13(5):1062–1084
14. van der Houwen PJ, Sommeijer BP, Couzy W (1992) Embedded diagonally implicit Runge-Kutta algorithms on parallel computer. Math. Comput. 58:135–159

Chapter 6

Comparative study of production control systems through simulation

K. Onan[1], B. Sennaroglu[2]

[1] Dogus University, Industrial Engineering Department, Acibadem
Kadikoy 34722 Istanbul TURKEY, konan@dogus.edu.tr
[2] Marmara University, Industrial Engineering Department, Goztepe
Kadikoy 34722 Istanbul TURKEY, sennaroglu@eng.marmara.edu.tr

Abstract. The purpose of this chapter is to simulate pull, push and hybrid production control systems for a single line, multi-stage and continuous production process, which is aluminium casting, in order to compare their performances where the current production control system is push system. This chapter also provides an approach in constructing hybrid system by using TSP in minimizing total setup time to obtain optimal sequence of orders.

Keywords. Production control systems, Pull system, Push system, Hybrid system, Simulation, Travelling salesman problem

6.1 Introduction

One of the most important issues for managers of manufacturing companies to decide on is what production control system would be the most appropriate for their companies. The choice is a matter of research and investigation but choosing the right system is a very important competitive advantage for the manufacturing companies.

Production systems are used to control the movement of product through the manufacturing process. A push system is defined by make to stock and a pull system is defined by make to order [1]. Materials resource

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_6,

planning is the best and most classical example to push systems, which uses past information to forecast the customer demands. Much of the discussion in the literature focuses on the relative merits of push (e.g. MRP) and pull (e.g. Kanban) systems [2]. In the case of a pull system the best example is Kanban approach. Although the just-in-time method is gaining popularity, the MRP philosophy is still quite compelling, as it not only incorporates the relationship between end items and components but also uses forecasts of future demand over a reasonable planning horizon. However, the two philosophies are not necessarily contradictory. Many manufacturing systems incorporate a hybrid of the two [3]. The relationship between MRP/JIT and push/pull strategies, major controversies and related literature for their comparison and integration is examined in detail by [4].

In the literature one can find studies in which simulation is used as a tool to evaluate and compare the performances of push and pull systems [5–7]. A hybrid push/pull production control algorithm is developed and tested for use in a multi-stage, multi-line, assembly-type repetitive manufacturing environment via simulation by [8].

The purpose of this study is to compare the performances of push, pull and hybrid production control systems for a single line, multi-stage and continuous process using simulation as a tool. The study is inspired by a production scheduling problem in a large aluminium rolling and processing factory in Istanbul. The production process is aluminium casting and the current production control system is a push system.

6.2 Production process

The production process takes place on a casting line and continuous production is required to avoid reheating a furnace after it cools down since it is a long and costly process. The first operation is the casting operation whereby ingots or slabs of pure aluminium are melted in a furnace. Then, additive elements such as magnesium, and vanadium are added to the furnace in specific amounts, which determine the alloy of the aluminium. The melted metal in the furnace flows through a shaper and a rolling mill (Fig. 6.1).

After the casting operation, to obtain the desired width and thickness, the metal is fed in coil form through a series of cold rolling mills, which successively reduce the metal thickness and recoil it after each rolling pass, getting it ready for the next until the required thickness is obtained. Annealing may be required between passes, depending on the final temper required. This is followed by surface processing and cutting operations which are called secondary operations. According to customer specifications the coils are rolled to the desired properties through these secondary operations.

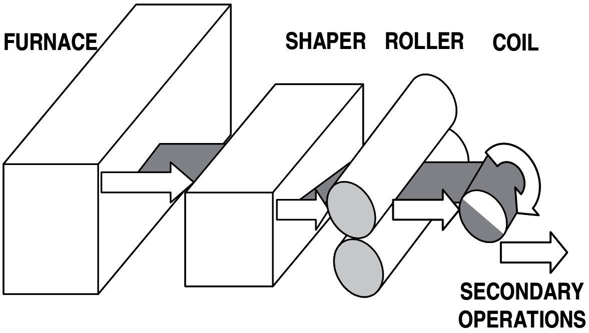


Fig. 6.1. Aluminium casting line

Final products can be used in the manufacturing of aircraft, satellites, space laboratory structures, tankers and freight wagons, buses, truck bodies, tankers, radiators, traffic signs and lighting columns, chemical process plants, chemical carriers, food handling and processing equipment, packaging, cans, bottle caps, wrapping, packs and containers.

The products ordered by the customers may have various properties. These properties are width, thickness and alloy. The important factors that need to be considered are listed below:

Width: If two consecutively scheduled jobs have different widths, then there will be a setup time after the first job. For two consecutive jobs, the setup time when the width of the second order is narrower than the first order is shorter than the setup time when the first order is narrower than the second order. This is because both shaper and roller must be changed to process a wider job.

Thickness: The orders may require different casting thicknesses. However, the change in thickness can be handled in a very short amount of time and minimally impacts setup in casting scheduling.

Alloy: If there is an alloy change between two consecutive orders, then a setup is required. Most of the time, no significant setup is required to process a more composite alloy after a purer alloy (i.e. some additive elements are added to the furnace and then production continues). However, in the reverse case, i.e. if a purer alloy is to be cast after a composite alloy, the furnace must be cleaned thoroughly (i.e. hot cleaning must be done), and this process usually takes significantly longer than the previous case.

Last Job of the Previous Schedule: Since the production is continuous, the current month's schedule should take into account properties of the last job of the previous schedule as the initial condition of width, thickness and alloy for minimizing total setup time.

6.3 Simulation models

This comparative study of production control systems consists of three simulation models. The first one is the simulation of the current system which is a push system. The second one is the simulation model of the pull system. The third one is the hybrid system which starts with a push system applied to the casting line and continues with a pull system applied to the secondary operations where the roller is the boundary between push and pull. The approach used in constructing hybrid model is that production orders for casting operation are obtained by applying MRP method with TSP optimization, then production goes on with Kanban signal for secondary operations.

The production control systems are modelled using ARENA simulation software which is a SIMAN-based simulation language. BestFit distribution fitting software is used to find the probability distributions to be used in the modules of the simulation models. The following assumptions are made:

- The system is a single line, multi-stage and continuous production system.
- Setup characteristics of the server modules are similar; they all follow the same distribution.
- Transportation time is negligible.
- Processing on all workstation is carried out without defects, a perfect quality conformance is assumed along the system.
- Each order being manufactured follows the same process routine.
- Production is assumed to be continuous, without any maintenance or failure, but the setup times are included into the probability distributions for times.

6.3.1 Push system model

Push system is the current production control system in the company. There is a master schedule of the production and this schedule is built with the MRP method. Orders are forecasted and the materials requirements are planned according to these forecasts. The information flow has the same direction as the material flow.

For push model, probability distribution of arrival time of the orders is included in the arrive module. The probability distributions of casting and secondary operations' processing times are included in the involved modules. These probability distributions are obtained from the original production orders data of the manufacturing system.

6.3.2 Pull system model

Schedule is being built according to customer demand in the pull system. The information flow has the opposite direction to the material flow. Production is of make to order type and hence inventory level is affected by the changes in customer demands.

For pull model, probability distribution of arrival time of the orders, which are obtained from the real customer demand data, is included in the arrive module. The module of secondary operations pulls the materials from the casting line module and casting operation module pulls the materials from the arrive module. Signal module is used to pull the materials from the preceding operation. Create, signal and dispose modules are used to create the first signal; after the first signal is created the model itself creates the following signals, which release the material being held in the wait module.

6.3.3 Hybrid system model

The simulation model of the hybrid system consists of two phases. During the first phase orders are being forecasted and these forecasts are turned into production plans with MRP method after which the master plan is revised by applying travelling salesman problem (TSP) the method to find out the production plan with minimum setup times and then the metal is cast according to this revised plan. Then these melted and cast-metal half-products are stored for secondary operations. After that these half products turn into finished products through customer demand, so this second phase is a pull system. Therefore, the whole model can be called a hybrid system. The production control system is built as an appropriate mixture of the push and pull systems. There is a boundary in this model, which separates the production system into two parts as push for casting and pull for secondary operations. This kind of a control strategy is the so-called hybrid. The first part of the production is scheduled according to the forecasts, and the outputs of this first part are pulled through the secondary operations according to the customer demand. In the push part of the model, information flow has the same direction as material flow while in the pull part, they are opposite.

For hybrid model, probability distribution of arrival time of the orders is included in the arrive module. Secondary operations pull the materials from the wait module and production is being pushed through this wait module. The wait module is the boundary between the two parts: push-oriented part and pull-oriented part. Signal module is used to pull the materials from the preceding operation. Create, signal and dispose modules are used to create the first signal and after the first signal is created the model itself creates the following signals, which release the material being held in the wait module.

6.3.4 Travelling salesman problem (TSP) approach

In the hybrid production control model TSP method was used to minimize setup times to obtain a better schedule. Probably the most famous routing problem is the travelling salesman problem, which can be described as follows: A travelling salesman is required to call at each town in his district before returning home. The salesman would like to schedule his visits so as to travel as little as possible. Thus, the salesman encounters the problem of finding a route that minimizes the total distance (or time or cost) needed to visit all the towns in the district [9].

The cities are considered to be the orders and the distances between the cities are considered to be the setup times between the orders. So the route acquired from the optimal solution of TSP is also the optimum sequence of the orders with the minimum total setup time.

Although the TSP is so simple to characterize, it is very difficult to solve. The TSP belongs to the class of NP-hard problems. Thus it is unlikely that any efficient algorithm will be developed to solve it. Because of its simplicity, however, the TSP has been one of the most studied problems in this class [9]. The following is the integer programming formulation of TSP:

N: number of cities

for $i \neq j$, c_{ij} = distance from city i to city j

c_{ii} = M (a very high distance compared to real distances)

x_{ij} is identified as 1 or 0 according to the conditions mentioned:

$x_{ij} = 1$, TSP solution offers to go from city i to city j

$x_{ij} = 0$, otherwise

TSP Formulation:

$$Min\sum_i \sum_j c_{ij}x_{ij} \tag{6.1}$$

$$\sum_i x_{ij} = 1 \text{ for every } j \tag{6.2}$$

$$\sum_j x_{ij} = 1 \text{ for every } i \tag{6.3}$$

$$u_i - u_j + N\,x_{ij} \le N-1 \tag{6.4}$$

$$\text{for every } i \ne j;\; i = 2,3,...,N;\; j = 2,3,...,N$$

$$\text{every } x_{ij} = 0 \text{ or } 1; \text{every } u_i \ge 0$$

The objective function is defined to minimize total setup time (Eq. (6.1)). The constraint group (Eq. (6.2)) makes the salesman to visit every city just once. The constraint group (Eq. (6.3)) makes the salesman leave every city just once. The constraint group (Eq. (6.4)) provides that any group of x_{ij}, which does not complete the tour, is not a possible solution and any group of x_{ij}, which completes the tour, is a possible solution [10].

In this study Lingo software is used to solve the TSP applied to the hybrid system model. This software finds the optimal solution for TSP by using branch & bound method with integer programming model.

6.4 Performance measures

The performance of the models is compared using the following performance measures:

- Total output: Total system output is measured as the average of 100 total runs which has the run length of 436,320 minutes each.
- Average number in queue: It is defined for all processing modules and measured as the average number of units waiting for each process.
- Utilization: It is defined as the busyness percentage and gives opinion about idle time.

6.5 Model validation

The current production system was a push-oriented system and after building the simulation model it was run. The results were compared to the production data of the current system and the model was validated. It was observed that from run length to the number of products manufactured, there is conformance between the model and the system in use (Table 6.1).

Table 6.1. Push-oriented model and current system in use

	Push system model	Current system in use
Total run length (min)	436,320	436,320
Total output (units)	58	60

6.6 Results

The results of the runs from simulation models are used to compare performances of push, pull and hybrid production control systems.

The results in Table 6.2 represent the comparison of three strategies according to the performance measures. The pull system seems to have the best performance according to the total output.

Table 6.2. Performance measures for each model

	Push	Pull	Hybrid
Total output	58	61	59
Avg. # in queue	0.61	0	1.68
Busyness (%)	82.26	100	81.60

Figure 6.2 is the graphic for comparing the total outputs of the models. As mentioned above pull seems to show the best performance. Also there are other parameters to be reviewed and the next one is average number of entities in casting queues (Fig. 6.3). As a property of the pull system the queues before the processing modules are eliminated so the value of the average number of entities in casting queue of pull model is zero. And also it seems that applying hybrid strategy to this production system causes an increase in the number of entities waiting for operation at the point of boundary (work in process). But still total output value of the hybrid model is better than the push model.

The last figure to be discussed is the busyness percentage of casting module. Again as a characteristic of the pull system, idle time seems to be eliminated (Fig. 6.4).

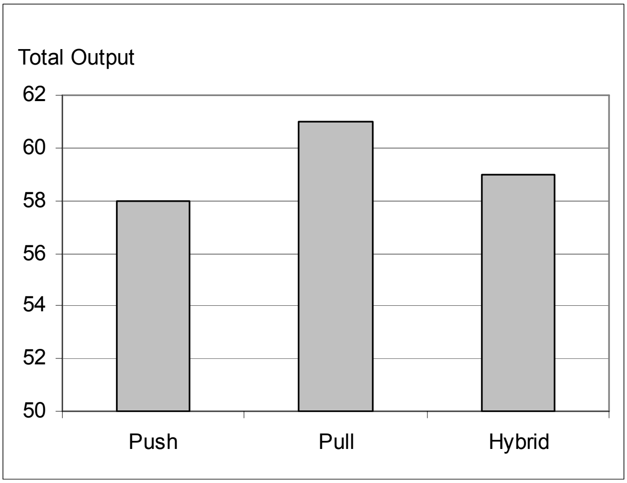


Fig. 6.2. Total output

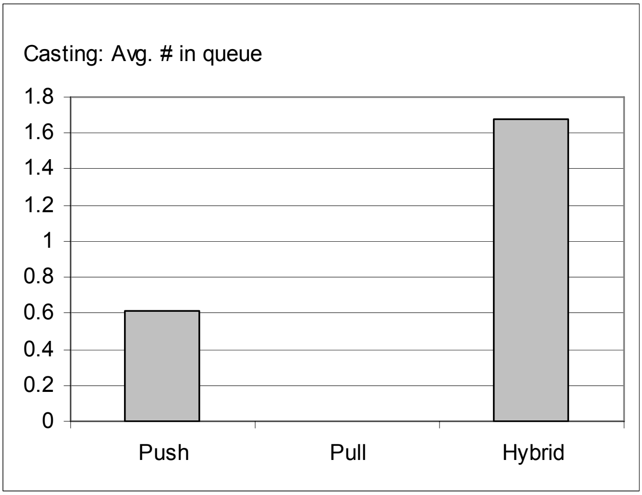


Fig. 6.3. Average number of entities in queue

These last two parameters show that the principal characteristics of pull system are observed on the model. So it can be said that the applications are valid, effective and successful. When these parameters are compared it is obvious that pull model seems to outperform the other models.

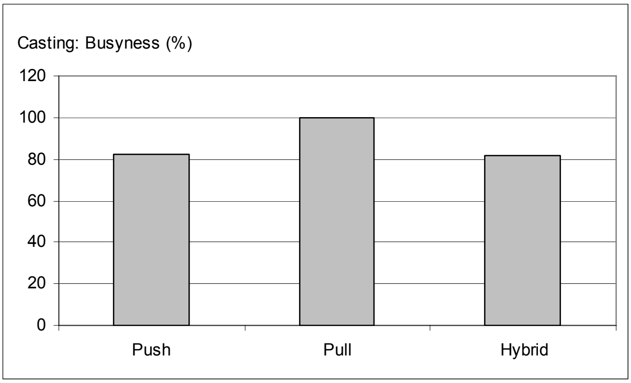
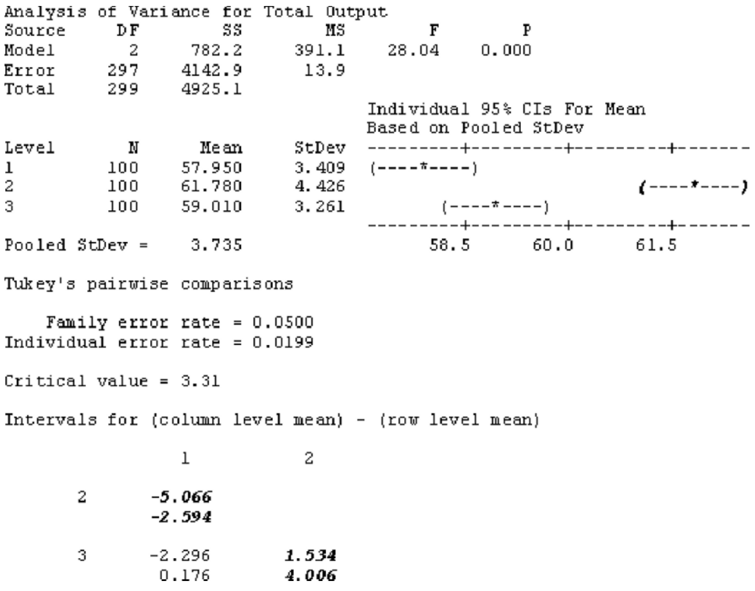


Fig. 6.4. Busyness percentage of casting line

To find out the significant differences of the models an ANOVA study was performed. Since in the ANOVA output (Fig. 6.5) p-value is less (p-value=0) than the level of significance (α=0.05) it is concluded that there are significant differences between average total outputs of models. Also Tukey's multiple comparisons procedure was used to indicate pairs which are significantly different and shows that pull is better than the others. The important parts and values of the figure are highlighted with bold and italic characters (Fig. 6.5).



```
Analysis of Variance for Total Output
Source     DF        SS        MS        F        P
Model       2     782.2     391.1    28.04    0.000
Error     297    4142.9      13.9
Total     299    4925.1
                                 Individual 95% CIs For Mean
                                 Based on Pooled StDev
Level       N      Mean     StDev  ---------+---------+---------+------
1         100    57.950     3.409  (----*----)
2         100    61.780     4.426                       (----*----)
3         100    59.010     3.261        (----*----)
                                   ---------+---------+---------+------
Pooled StDev =    3.735                  58.5      60.0      61.5

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0199

Critical value = 3.31

Intervals for (column level mean) - (row level mean)

                 1          2

       2      -5.066
              -2.594

       3      -2.296      1.534
               0.176      4.006
```

Fig. 6.5. ANOVA and Tukey's multiple comparison results

6.7 Conclusion

Production control systems are one of the most important issues for managers to decide. The choice between different systems is a matter of research and investigation. But choosing the most appropriate system is a very important competitive advantage for the manufacturing companies. Here, in this study, using simulation the most appropriate production control system was searched among push, pull and hybrid systems for a multi-stage single-line continuous production system. Arena simulation software was chosen as the simulation tool and BestFit distribution fitter software was chosen for the statistics studies such as determining the probability distribution of the data. Results of the three systems were compared for the performance measures through graphs. The performance measures are total output, average number in queue and utilization. It was clear that the pull system was the best and push system was the worst choice as a production control system. But before a manufacturing organization can enjoy the benefits of pull, it must be aware of the fact that the lean manufacturing philosophy must be accepted and this may require it to change or modify its operating and management procedures; even it may have to make adjustments for its plant layout. Therefore, maybe it is better for a manufacturing organization to adopt hybrid production control system for avoiding vital changes required by pull system. As this study indicates that hybrid system outperforms push system by using both MRP and Kanban and also the TSP method for minimizing total setup time, this may cause less sufferance for the manufacturing organization.

References

1. Lubben RT (1988) Just in time manufacturing: An aggressive manufacturing strategy. McGraw-Hill, New York
2. Spearman ML, Zazanis MA (1992) Push and pull production systems: Issues and comparisons. Operations Research 40:521–532
3. Nahmias S (1993) Production and operations analysis. 2nd edn. Irwin, Illinois
4. Benton WC, Shin H (1998) Manufacturing planning and control: The evolution of MRP and JIT integration. European Journal of Operational Research 110:411–440
5. Thesen A (1999) Some simple, but efficient, push and pull heuristics for production sequencing for certain flexible manufacturing systems. International Journal of Production Research 37(7):1525–1539

6. Kim K, Chhajed D, Palekar US (2002) A comparative study of the performance of push and pull systems in the presence of emergency orders. International Journal of Production Research 40(7):1627–1646
7. Weitzman R, Rabinowitz G (2003) Sensitivity of push and pull strategies to information updating rate. International Journal of Production Research 41(9):2057–2074
8. Beamon BM, Bermudo JM (2000) A hybrid push/pull control algorithm for multi-stage, multi-line production systems. Production Planning and Control 11:349–356
9. Evans JR, Minieka E (1992) Optimization algorithms for networks and graphs. 2nd edn. Marcel Dekker, New York
10. Winston WL (2004) Operations research. 4th edn. Thomson – Brooks/Cole, Belmont

Chapter 7

A DEA and goal programming approach to demand assignment problem

Y. Ekinci

Department of Industrial Engineering, Dogus University, Zeamet S. No:21
Acibadem/Istanbul TURKEY, yekinci@dogus.edu.tr

Abstract. Assignment problem has always been a popular problem for production and operations research. Many methods and heuristics have been developed for these kinds of problems. This research deals with the problem of an automotive company's driver belt assembly supply from its by-industry suppliers. First, the suppliers are evaluated by data envelopment analysis (DEA), which is a mathematical modelling approach finding the relative efficiencies of the decision-making units (DMUs). Later on, the suppliers that are found inefficient are eliminated and only the efficient suppliers are included for the assignment problem. Preemptive goal programming (PGP) is applied for the assignment of demand in order to satisfy the goals and the demand of the company by taking into account the priorities given by the company.

Keywords. Data envelopment analysis (DEA), Preemptive goal programming, Demand assignment

7.1 Introduction

High competition reality in today's world forces companies to work with lower costs, higher efficiencies and higher qualities. This reality is also the reason for the companies' cooperation with suppliers which help them to achieve these goals. Selecting the suppliers satisfying this requirement and

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_7,

assigning them the demands lead to assignment problems for the companies. Especially, in the automotive industry, where just-in-time manufacturing is performed, supply problems have a great importance. This chapter proposes a two-step solution to the demand assignment problem of an automotive company for driver belt assemblies.

Assignment problem has always been a popular problem for production and operations research. Many methods and heuristics for these kinds of problems have been developed. These problems usually include a company, which is trying to assign the demand to the suppliers, and therefore some suppliers, criteria, goals and constraints. In order to solve the assignment problems, transportation methods can also be used such as northwest corner, minimum cost and vogel methods. The basic assignment method is the Hungarian method, which can be efficiently used for solving $m*m$ assignment problems [1]. But these are used for the cases where there is only one objective, usually cost, distance minimization or profit maximization. If there are many complex constraints and objectives, superior methods are necessary. Actually, once the problem is modelled mathematically, many operations research solving techniques can be applied. If there are more than one goal, then goal programming model which is one of the mathematical modelling applications can be of use. Since these problems also include selecting the right supplier, considering some criteria, multicriteria decision-making (MCDM) methods are also involved in the decision process. There are so many MCDM techniques such as analytical hierarchy process (AHP), TOPSIS (technique for order preference by similarity to ideal solution), ELECTRE I, II, III (elimination and choice translating reality english), PROMETHEE I,II. When data envelopment analysis, which is the technique applied in this research, is compared to other MCDM methods, it is seen that it requires less information from decision makers and analysts and it provides ranked alternative valuations that may be useful for some decision makers [2]. MCDM techniques are not used in the decision process in research, because the main idea of the decision is collaborating with the efficient suppliers, which will minimize their input values. Therefore data envelopment analysis (DEA), which has been employed successfully for assessing the relative performance of a set of firms, usually called decision-making units (DMU), which use the same inputs to produce the same outputs [3], is applied to the input values of suppliers in order to evaluate their relative efficiencies which will help the decision of supplier selection before moving to the assignment problem. After selecting the efficient suppliers by DEA, preemptive goal programming model is used to solve the multi-objective assignment problem of the automotive company.

The composition of this chapter includes literature review, first, on supplier evaluation and second DEA and, second, on preemptive goal programming. After literature review, problem definition and application of the models and their results are presented. Finally results are discussed and conclusion and further research ends this chapter.

7.2 Literature review

7.2.1 Supplier evaluation and DEA

Modern supply chain management, which is defined as the coordination and integration of the products and information from raw material to the final customer, emphasizes the importance of purchasing; additionally implementation of just-in-time production results in the analysis of purchasing management once again [4]. It has been reported that a majority of quality problems of an organization are due to defective material, and carefully selected, competitive suppliers can go a long way in minimizing adverse impacts and in fact in enhancing positive impacts on the quality of output of an organization, which leads to the fact that supplier selection is a crucial part of the functioning of an organization [3]. Weber and some other researchers scanned many articles on supplier evaluation in 1991 and 1993 and investigated the effect of JIT strategy over supplier evaluation using Dickson's 23 criteria; among 74 articles, net price, delivery and quality were discussed mostly with the percentage of 80, 59 and 54, respectively [5]. Therefore, these three criteria are used as inputs for the supplier selection part of the problem, also adding the reputation criteria, which has a high importance for the company, applying data envelopment analysis (DEA), which is the first step.

DEA is able to measure multiple inputs and outputs, which means it can operate as a multi-criteria decision-making (MCDM) tool, but DEA does not require assigned numeric weights or modelling preferences for analysis, although these could be introduced if/when desired [2]. This advantage and the objectivity property of DEA with the efficiency calculations is the main reason for using DEA for supplier selection which is the first step of the problem solution. Charnes et al. [6] first proposed DEA as a generalization of the framework of Farrell [7] on the measurement of productive efficiency [8,9]. DEA is a method for mathematically comparing different decision-making units' (DMUs) productivity, based on multiple inputs and outputs [8]. It is based on the economic notion of Pareto optimality: a given DMU is not efficient if some other DMU or some combination of other DMUs can produce the same amounts of outputs with less of some resources and not more of any other [9]. Multiple input, single output DEA model, utilizing the Pareto-Koopmans efficiency measure, will be used in this study with reference to Weber [10,11].

7.2.2 Preemptive goal programming

Goal programming (GP) which is a multi-objective decision-making method was first proposed by Charnes and Cooper in 1961. It was improved by Lee in 1972 and by Ignizio in 1976 [12]. Nowadays GP is acknowledged as one of the most effective strategies used in multi-objective optimization problems [13]. GP attempts to reach all the goals at the same time, which may affect each other negatively. When one goal reaches the targeted value, another one may move away from its own target, therefore objective function in GP models is usually to minimize the deficiencies from the aimed values. Some priorities can be put for the goals, so that the deficiencies that should be minimized have a sequence to be satisfied; this type of goal programming models are called preemptive goal programming models. In an initial step a first part model is solved which only incorporates the first-priority goals; if the execution of the initial step leads to more than one optimal solution of the first part model, a second part model incorporating the second-priority goals is solved keeping the optimal achievement level of first-priority goals constant; lower-priority goals are not considered unless the higher-priority goals are optimally satisfied and this optimal solution is many-valued [14]. Preemptive goal programming is used in the second step of the problem, which tries to find the optimal assignment of the demand to the efficient suppliers, found in the first step by DEA, satisfying the constraints and goals of the company.

7.3 Problem definition and application of the model

This chapter proposes a two-step solution to the demand assignment problem of an automotive company for driver belt assemblies. The company is trying to make the assignment of the demand to the suppliers while satisfying some goals and constraints. There are five suppliers that the company works with. Table 7.1 shows the criteria values, which are determined by the literature review and company experts, as price, quality (rejection rate), delivery (late delivery rate) and reputation. Price is measured in USDs, rejection and delivery rates are measured as percentages of total materials supplied by that supplier – which are kept in the database of the company – and reputation is measured in a 1–5 scale where 1 denotes the highest reputation. The first step is to find the efficient suppliers by using DEA and the second step is to assign the demand to the efficient suppliers by preemptive goal programming.

Table 7.1. Input values of the suppliers

	Supplier 1	Supplier 2	Supplier 3	Supplier 4	Supplier 5
Price ($)	20	22	23	23	22
Rejection (%)	1.2	1.5	1	1.5	1.3
Late delivery (%)	2	7	2	7	3
Reputation	1	2	2	3	2

Multiple input, single output DEA model, utilizing the Pareto-Koopmans efficiency measure, will be used in this study's first step, referencing to Weber [10]. This form of the model measures the efficiency of DMUs by how well they minimize multiple input criteria to produce a single unit of output [11]:

$$\text{Min}\, x_k - \varepsilon\left(\sum_{i=1}^{m} s_i\right) \tag{7.1}$$

subject to

$$x_k \cdot w_{ik} - \sum_{j=1}^{n} w_{ij} \cdot y_j - s_i = 0 \text{ for all } i\text{=1,…,}m \tag{7.2}$$

$$\sum_{j=1}^{n} y_j = 1 \tag{7.3}$$

$$y_j \geq 0 \text{ for all } j = 1,\ldots,n \tag{7.4}$$

$$s_i \geq 0 \text{ for all i} = 1,\ldots,m \tag{7.5}$$

$$x_k \text{ unconstrained but assumed positive} \tag{7.6}$$

where

x_k is the Farrell's efficiency measure for supplier k,
s_i are input criteria slack variables,
y_j are reference weights associated with vendor j,
ε is an infinitely small number,
w_{ij} is the input criteria value for the ith criteria and the jth supplier,
m is the number of criteria, and
n is the number of suppliers.

The model above is written for each supplier *k* using the data in Table 7.1. The *m* and n values are as follows: $n = 5$, there are five suppliers; $m = 4$, there are four inputs (price, rejection, late delivery and reputation). The efficiencies derived after solving all of the five models are seen in Table 7.2.

Table 7.2. Supplier efficiencies

Supplier	Efficiency
Supplier 1	1.000
Supplier 2	0.909
Supplier 3	1.000
Supplier 4	0.870
Supplier 5	0.916

Supplier 1 and Supplier 3 are relatively efficient in the given data set compared to other suppliers as their efficiency values are 1.000. Actually, the efficiency values of the other suppliers are not very low; Supplier 2 has the efficiency value 0.909, Supplier 4 has the efficiency value 0.870 and Supplier 5 has the efficiency value 0.916.

For the second step of the problem, a preemptive goal programming model is formed and solved. There are two functional constraints of the company. One is about satisfying the demand and the other is about the budget. Demand of the company is 10.000 pieces per month (see Eq. (7.8)). Budget constraint is \$210.000 per month (see Eq. (7.9)). There are three goal constraints of the company. The goal about the quality has the first priority, and the company does not want the rejection rate to be more than 1.1% (see Eq. (7.10)). The second goal of the firm is keeping the late delivery rate under 1.5% (see Eq. (7.11)). The last goal of the company is about the reputation of the suppliers and its preference is working with suppliers having reputation degree of 1 and 2 over 5 (see Eq. (7.12)). Equation. (7.7) is the objective function of the preemptive goal programming model.

x_i = Number of driver belt assemblies bought from supplier *i*; *i* = 1,2 (1: Supplier 1, 2: Supplier 3)

$$\text{Min } P_1 s_1^+ ,\, P_2 s_2^+ ,\, P_3 s_3^+ \tag{7.7}$$

subject to

$$x_1 + x_2 \geq 10.000 \tag{7.8}$$

$$20x_1 + 23x_2 \leq 210.000 \tag{7.9}$$

$$1.2\,x_1 + 1\,x_2 + s_1^- - s_1^+ = 11.000 \tag{7.10}$$

$$2x_1 + 2x_2 + s_2^- - s_2^+ = 15.000 \tag{7.11}$$

$$1x_1 + 2x_2 + s_3^- - s_3^+ = 20.000 \tag{7.12}$$

$$x_1, x_2, s_1^-, s_1^+, s_2^-, s_2^+, s_3^-, s_3^+ \geq 0 \tag{7.13}$$

After solving the goal model it is found that 6667 belt assemblies should be bought from Supplier 1 and 3333 belt assemblies should be bought from Supplier 2. This result leads to positive deficiency of 333 from the first goal and positive deficiency of 5000 from the second goal and there is no positive deficiency from the third goal, therefore the objective function results in 5333.

7.4 Conclusion

Assignment problem has always been popular for production and operations research. Many methods and heuristics have been developed for these kinds of problems. This chapter represents a two-step solution for the assignment problem of an automotive company for purchasing driver belt assemblies where these problems have a great importance because of the just-in-time manufacturing. First, the five suppliers of the material are evaluated by data envelopment analysis (DEA), which is a mathematical modelling approach finding the relative efficiencies of the decision-making units (DMUs).

The criteria are determined by the literature review and company experts as price, quality (rejection rate), delivery (late delivery rate) and reputation. Multiple input, single output DEA model, utilizing the Pareto-Koopmans efficiency measure, is used for supplier evaluation. Later on, the suppliers that are found inefficient are eliminated and only the two efficient suppliers are included for the assignment problem. Preemptive goal programming (PGP) is applied for the assignment of demand in order to satisfy the goals of the criteria by taking into account the priorities given by the company and the constraints about budget and demand constraints. The solution of the goal model assigns the demand to the suppliers.

The two-step solution used in this chapter prevents the company from making the assignment among all suppliers and allows the company to make the assignment among only the efficient suppliers and DEA is an effective way of comparing the efficiencies of units.

References

1. Winston WL, Venkataramanan M (2005) Introduction to mathematical programming. Thomson Learning Academic Resource Center, Pacific Grove, pp 405–406
2. Wong WP, Wong KY (2007) Supply chain performance measurement system using DEA modeling. Industrial Management & Data Systems 107(3):361–381
3. Ramanathan R (2007) Supplier selection problem: integrating DEA with the approaches of total cost of ownership and AHP. Supply Chain Management: An International Journal 12(4):258–261
4. Seydel J (2006) Data envelopment analysis for decision support. Industrial Management & Data Systems 106(1):81–95
5. Franklin Liu FH, Hui LH (2005) The voting analytic hierarchy process method for selecting supplier. International Journal of Production Economics 97:308–317
6. Charnes et al. (1978)
7. Farrell (1957)
8. Donthu N, Hershberger EK, Osmonbekov T (2005) Benchmarking marketing productivity using data envelopment analysis. Journal of Business Research 58:1474–1482
9. Al-Shammari M (1999) Optimization modeling for estimating and enhancing relative efficiency with application to industrial companies. European Journal of Operational Research 115:488–496
10. Weber (1996)
11. Weber CA (2006) A data envelopment analysis approach to measuring vendor performance. Supply Chain Management 1(1):28–39
12. Bal H, Örkcü HH, Çelebioğlu S (2006) An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems. Computers & Industrial Engineering 50:296–311
13. Abdelaziz FB (2007) Multiple objective programming and goal programming: New trends and applications. European Journal of Operational Research 177:1520–1522
14. Peters ML, Zelewski S (2007) Assignment of employees to workplaces under consideration of employee competences and preferences. Management Research News 30(2):84–99

Chapter 8

Space–time mixture model of infant mortality in peninsular Malaysia from 1990 to 2000

N. Abdul Rahman[1], A.A. Jemain[2]

[1] School of Mathematical Sciences, Faculty of Science & Technology,
Universiti Kebangsaan Malaysia, 43000 Bangi, Selangor,
Malaysia, nuzlinda2001@yahoo.com
[2] School of Mathematical Sciences, Faculty of Science & Technology,
Universiti Kebangsaan Malaysia, 43000 Bangi, Selangor,
Malaysia, azizj@pkrisc.ukm.my

Abstract. Disease mapping is a method used to display the geographical distribution of disease occurrence. Some traditional methods of classification for detection of high- or low-risk area such as traditional percentiles method and significant method have been used in disease mapping for map construction. However, as described by several authors, the classification based on these traditional methods has some disadvantages for describing the spatial distribution of the risk of the disease concerned. To overcome these limitations, an approach using space–time mixture model within an empirical Bayes framework is described in this chapter. The aim of this chapter is to investigate the geographical distribution of infant mortality in peninsular Malaysia from 1991 to 2000. The analysis showed that in the early 1990s the spatial heterogeneity effect was more prominent; however, toward the end of 1990s this pattern tends to disappear. Indirectly, this may indicate that the provisions of health services throughout peninsular Malaysia are uniformly distributed over the period of the study, particularly toward the year 2000.

Keywords. Disease mapping, Space–time mixture model, Infant mortality, Geographical distribution, Spatial heterogeneity

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_8,

8.1 Introduction

Child health is a central issue amongst the public in many developing countries. Infant mortality rate is one of the most common measure used to describe the level of services relating to health, socio-economic, and education of a country. Since independence in 1957, Malaysia had experienced a very remarkable decline in infant mortality from the rate of around 100 per 1000 to around 13 per 1000 by the late 1980s. It was reported that this rate has been reduced to 9 per 1000 in 2004. This achievement is nearly equal to the rate experienced by developing countries such as United States and Britain, with 7 and 6 deaths per 1000, respectively. The decline in infant mortality rate in Malaysia could possibly be due to the prosperous socio-economic situation where the average incomes have increased over the years. Moreover, basic facilities such as water supply, electricity, sewage, sanitation, and health services have being improved, provided to the wider population of the country. Apart from that, the levels of education and health consciousness have increased among Malaysians and so too have other factors that directly and indirectly influence the infant mortality in Malaysia such as ethnicity, mother's education, preceding birth interval, and birth place [13].

The basic concept of mapping is to group the information in the data for all regions in the study area into several components or exclusive groups, where individual region in the same component has a similar risk. One way of displaying the variability of disease or mortality rate is by a widely used technique called disease mapping. It is very useful to produce such maps especially for government agencies in resource allocation or identifying hazards that contribute to the disease [9]. Usually, in the context of health sector, the authority in charge aims to identify whether the risks for a particular disease concerned are uniformly distributed or homogeneous for different regions of the country. For example, as mentioned earlier, it is fortunate that the infant mortality rate in Malaysia has improved over the last few decades, but the issue concerned is whether the improvement is uniformly distributed throughout the country. Does every district experience the same level of improvement or reduction of the risks? Does the improvement only occur in certain areas while the other areas still remain in the high-risk area category? If there is a huge gap between the high-risk areas and the low-risk areas, the disease risks can be divided in to several categories or considered as heterogeneous. This is the main issue that will be addressed in this chapter in the context of disease mapping of infant mortality in Malaysia.

In disease mapping, let us divide the study area to be mapped into M mutually exclusive districts $(i=1,2,\ldots,M)$. Each district has its own observed number of cases, o_i, and expected number of cases, E_i. The expected number of cases is calculated as follows:

$$E_i = N_i \sum o_i \Big/ \sum N_i \tag{8.1}$$

where N_i is the population for area i [7]. Here the standardization is done on the total population at risk. The standardization can be done on other factors such as age and gender, and this method has been discussed by several authors [10,14].

It is common to assume that the observed number of cases, o_i, follows a Poisson distribution with expectation $E_i\theta$ and the probability density function is defined as

$$\Pr(O_i = o_i) = \frac{\exp(-\theta E_i)(\theta E_i)^{o_i}}{o_i!} = f(o_i, \theta, E_i) \tag{8.2}$$

where θ is the relative risk of disease concerned over the study area. Using o_i and E_i as obtained based on the data, we can have one of the most common index to estimate the relative risk for region i, θ_i, i.e., Standardized mortality ratio (*SMR*) defined as

$$\hat{\theta}_i = SMR_i = \frac{o_i}{E_i} \tag{8.3}$$

In map construction, the important elements are obtaining smoothed estimators of relative risk and categorizing or classifying of all districts into several components using shading or coloring to differentiate the level of risks for each component. Although *SMR* has been used commonly as an index to measure relative risk, however, it has some weaknesses where the variance of *SMR*, θ_i / E_i, depends on E_i. The variance will be large when the expected value is small, as contributed by the small population size, and the variance will be small when the expected value is large due to the large population size. If the observed value is zero such as in the case of rare disease, the *SMR* and the standard deviation will be zero. Another limitation of *SMR* is the instability of the relative risk estimation due to the presence of extreme *SMR* when rare diseases are investigated in small population areas [15].

To overcome the drawbacks of the *SMR* a Bayesian approach has been used which allows for the risk to vary between the different districts as given by the assumption

$$O_i \sim Poisson\left(E_i\theta_i\right) \tag{8.4}$$

and the probability density function is defined as

$$\Pr(O_i = o_i) = \frac{\exp(-\theta_i E_i)(\theta_i E_i)^{o_i}}{o_i!} = f(o_i, \theta_i, E_i) \tag{8.5}$$

For example, the empirical Bayes of the relative risks where a random effects (or mixture) model that assumes a parametric probability density function (pdf), denoted as $f(\theta)$, for the distribution of relative risks between districts was adopted [16]. This modeling approach has been used in many fields including disease data by applying several empirical Bayes methods of estimation to smooth the *SMR* [3,11,12]. Some authors have provided discussion on hierarchical Bayesian approach with structured and unstructured spatial random effects [8]. Although the Bayesian methods can provide estimate on relative risks for each district, the number of optimum classification for categorizing the districts cannot be obtained based on them. The most common approach that is widely used by many researchers for categorizing areas in disease mapping is classification based on quartiles. However, this method is rather arbitrary and has no guarantee in detecting the classification of high-or low-risk areas. Another disadvantage of the Bayesian relative risks estimation is the usage of assumption in Eq. (8.4), which will give the number of parameters and the number of districts as the same. If the number of districts is large, there might be difficulties in estimating parameters consistently because of too many parameters to be estimated. As an alternative approach, a method has been suggested to overcome these drawbacks which include the time factor and consider spatial heterogeneity effect will be discussed in this chapter known as space–time mixture model within non-parametric approach for map construction. This method has appeared to be very attractive for practical applications and has become a more flexible tool [2]. Infant mortality data in peninsular Malaysia from 1991 to 2000 will be applied using this approach to examine the geographical distribution of the disease concerned.

8.2 Methodology

Space–time mixture model is an extension model of mixture model by including the time factor in order to study the disease pattern in certain period of time. This model gives a valuable indication of an emerging pattern over time because it looks simultaneously for all space–time components [15]. The basic idea of space–time mixture model approach in the context of disease mapping is to consider all space–time data as a single data set. The same steps of modeling the mixture model will be applied in estimating parameters, so in this chapter, the discussion on mixture model will be presented in application to space–time data. In mixture model, we assume that the population comes from several heterogeneous components where every component consists of different risk levels of disease. This assumption will give a more heterogeneous case. Assume that the mixture model consists of c components and each component has a disease risk θ_j, where $j=1,\ldots,c$. Let p_j denote the proportion of regional areas having θ_j risk. This discrete parameter distribution P for describing the level of risk can be given as

$$P=\begin{bmatrix}\theta_1,\ldots,\theta_c\\ p_1,\ldots,p_c\end{bmatrix} \tag{8.6}$$

Accordingly, we may assume that the observed data in district i of a particular year t, o_{it}, comes from a non-parametric mixture density identified in the following form:

$$f\left(o_{it},P,E_{it}\right)=\sum_{t=1}^{T}\sum_{j=1}^{c}p_j f\left(o_{it},\theta_j,E_{it}\right) \tag{8.7}$$

where $p_1+\cdots+p_c=1$, $p_j\geq 0, j=1,\ldots,c$ and $t=1,\ldots,T$. E_{it} is the expected number of cases in district i of a particular year t. The number of parameters to be estimated in the model with c components considered above are $2c-1$ which consists of c unknown relative risks $\theta_1,\ldots,\theta_c$ and $c-1$ unknown mixing weights $p_1,\ldots,p_{c-1}$ where $f\left(\cdot\right)$ denotes the Poisson density taken from previous assumption [6].

One of the basic issue in mixture model is whether the number of components, c, is unknown or assumed to be known [18]. They called these two cases as flexible support size and fixed support size, respectively. However, in both cases, the maximum likelihood approach can be applied for the parameter estimation. In the estimation based on flexible support size, a grid containing $\theta_j s$ is defined and the corresponding p_j that maximized the likelihood function is determined. However, in this chapter, the fixed support size is considered and the algorithms used for this estimation is the EM algorithm [19].

In EM algorithm, the first step of mixture model involves estimating θ_j and p_j in each component by giving their initial values. These initial values and the number of components to be estimated can be obtained from the histogram of relative risks or *SMR* where the height of the bars may be used as the estimate for proportion corresponding to relative risks while the number of bars may be used as the number of components. We are interested in determining the membership of each district to a particular component. Let us denote the full data as $\left(o_{it}, E_{it}, x_{i1t}, x_{i2t}, \ldots, x_{ict}\right)$ where x_{ijt} indicate the membership of district in the th component for the year *t*. For example, if region *i* in the year *t* belongs to the third component, it can be written as $x_{it} = \left(0,0,1,0,\ldots,0\right)^T$. Based on the information of the initial weights $\hat{p}_j$ and relative risks $\hat{\theta}_j$ obtained, we can execute the EM algorithm which consists of E-step and M-step. The E-step consists of the calculation of the probability of each district belonging to *j*th component while the M-step consists of the calculation of the weights and relative risks. These two steps will be repeated alternately until the convergence criterion is met and can be summarized as follows:

E-Step

$$w_{ijt}{}^{(r)} = \Pr\left(X_{ijt} = 1 \middle| o_{it}, P, E_{it}\right) = \frac{\hat{p}_j{}^{(r)} f\left(o_{it}, \theta_j{}^{(r)}, E_{it}\right)}{\sum_{j=1}^{c} \hat{p}_j{}^{(r)} f\left(o_{it}, \theta_j{}^{(r)}, E_{it}\right)} \tag{8.8}$$

M-Step

$$\hat{p}_j{}^{(r+1)} = \frac{\sum_{i=1}^{M} w_{ijt}{}^{(r)}}{M} \text{ and } \theta_j{}^{(r+1)} = \frac{\sum_{i=1}^{M} w_{ijt}{}^{(r)} \frac{o_{it}}{E_{it}}}{\sum_{i=1}^{M} w_{ijt}{}^{(r)}} \tag{8.9}$$

When the convergence is obtained, the next step is to compute the non-parametric maximum likelihood estimator (NPMLE) that maximizes the log-likelihood function which is defined as

$$l_c = \sum_{t=1}^{T}\sum_{i=1}^{M} \log f\left(o_{it}, P, E_{it}\right) = \sum_{t=1}^{T}\sum_{i=1}^{M} \log\left\{\sum_{j=1}^{c} p_j f\left(o_{it}, \theta_j, E_{it}\right)\right\} \tag{8.10}$$

Further step is to determine the most suitable number of components by computing the difference between the log-likelihood for *c* components and $c+1$ components, which is known as likelihood ratio statistics (*LRS*) and is defined as follows:

$$LRS = -2\left(l_c - l_{c+1}\right) = -2\log\theta \tag{8.11}$$

The purpose of calculating the *LRS* is to test this hypothesis:
H_o : number of components is *c*
H_a : number of components is $c+1$

A problem arises in determining the number of components when the solution consists of the log-likelihood values that are nearly the same for every component. Conventionally, the LRS test has an asymptotic chi-square distribution with degrees of freedom equal to the difference between the number of parameters under the alternative and null hypotheses. However, these conventional results for LRS do not hold for mixture, and a method proposed to obtain the critical values in determining the number of components is via a simulation technique, for example by parametric bootstrap [18].

Once the optimum number of components is obtained, the final step in mixture model approach is to classify the membership of each district to a component. Classification can be obtained by applying Bayes' theorem which involves computing the probability of each district belonging to each component with the posterior probability given by

$$\Pr\left(X_{ijt} = 1 \middle| o_{it}, P, E_{it}\right) = \frac{\hat{p}_j f\left(o_{it}, \theta_j, E_{it}\right)}{\sum_{j=1}^{c} \hat{p}_j f\left(o_{it}, \theta_j, E_{it}\right)} \tag{8.12}$$

where $i = 1, \ldots, M, j = 1, \ldots, c,$ and $t = 1, \ldots, T$. The *i*th district in the year *t* will belong to the component or subpopulation *j* if the posterior probability of this belonging is highest.

8.3 Result

Data analysis based on space–time mixture model is illustrated using the infant mortality data in peninsular Malaysia from the year 1991 to 2000. Comparisons of the results obtained by this method throughout the study period become easier as all maps for each particular year have the same categorization. Table 8.1 below shows the result on how to determine the optimum number of components based on log-likelihood, l_c, mentioned above.

From this table, the models with four and five components have the lowest log-likelihood value, which is the same. Since there is no improvement in log-likelihood value for space–time mixture model with five components and by considering the parsimony factor, we choose a model with four components as the best model to fit the space–time data used in this study. Although it has been suggested by some studies that bootstrap method should be applied in deciding to choose either c or $c+1$ components, we based our decision on the previous argument. From the fitted model with four components, the first category had the lowest risk with mean of 0.726 and weight of 0.324 and the highest risk category with mean of 2.199 and weight of 0.014. The analysis of the space–time infant mortality data in peninsular Malaysia over the last decade leads to the mixture density with four components given by

$$\begin{aligned} f\left(o_{it},\hat{P},J_{it}\right) &= f\left(o_{it},0.726,J_{it}\right)\times 0.324+f\left(o_{it},1.131,J_{it}\right)\times 0.550+ \\ & f\left(o_{it},1.630,J_{it}\right)\times 0.112+f\left(o_{it},2.199,J_{it}\right)\times 0.014 \end{aligned} \tag{8.13}$$

Corresponding to the results given in the table, we can summarize the geographical distribution of infant mortality throughout the study period as given in Figs. 8.1, 8.2, and 8.3 by providing the space–time maps for the year 1991, 1996, and 2000. As each map obtained throughout the study period has the same classification, it is easier to compare and interpret the fluctuation of the disease concerned over time. In the early 1990s, it can be seen that only about 6% of the districts fall in lowest risk areas; however, in the middle and late 1990s, the infant mortality had improved with almost 50% or more of the districts belonging to this category. This figure also shows that none of the districts were in the highest risk category in 1996 and 2000 but four districts were in this category in 1991. These changes indicate that infant mortality in peninsular Malaysia had improved over the last decade and tends to be more homogeneous toward the end of the study period.

Table 8.1. Result of space–time mixture model

Number of components (c)	Mean relative risks $\left(\hat{\theta}_j\right)$	Weight ($\hat{p}_j$)	Log-likelihood (l_c)	LRS
5	0.726	0.324	2625.094	0.000
	1.131	0.550		
	1.630	0.112		
	2.199	0.012		
	2.199	0.002		
4	0.726	0.324	2625.094	72.420
	1.131	0.550		
	1.630	0.112		
	2.199	0.014		
3	0.749	0.370	2661.304	330.504
	1.175	0.547		
	1.810	0.083		
2	0.834	0.564	2826.556	1515.370
	1.374	0.436		
1	1.070	1.000	3584.241	–

LRS likelihood ratio statistics.

8.4 Discussion

For quite some time, many researchers have conducted various studies in disease mapping using the traditional methods of classification such as percentiles method and significant method. However, these methods have some deficiencies and potential of misrepresenting the graphical distribution, and questions regarding whether these classifications give a correct interpretation may be raised [18]. An alternative approach suggested is the mixture model that could produce a smoother map where the random variability has been extracted from the data. Other main advantages of using the mixture distribution are its discreteness, straight forward map construction, and providing the optimum number of components. The inclusion of space–time factor in mixture model satisfactorily produces maps that are easier to interpret and compare by looking at the maps separately since maps for each year throughout the study period have the same classification.

Based on the three maps in Figs. 8.1, 8.2, and 8.3, it is very clear that the space–time mixture model has removed random variability from the map and provides a better and clearer picture of classification for high-and low-risk areas. For the period of 10 years, i.e., from 1991 to 2000, we can conclude that the classifications tend to be more homogeneous implying that the random variability has reduced with time. Furthermore, toward the end of the study period, maps obtained showed that more districts have fallen into the low–risk categories which indicate that infant mortality in peninsular Malaysia has improved within the last decade.

There are many factors that contribute to the reduction of infant mortality. Some literatures stated that infant mortality is more likely to be related to the socio-economic level, health behavior, quality of antenatal care, support during delivery, postnatal care, nutritional status, education level, unemployment, and birth intervals [1,17,21]. These factors were addressed in the case of Malaysia as shown by the increase in health of RM17.30 per capita in 1970 to RM248 per capita in the year 2000 [4]. The number of hospitals increased from 84 public hospitals in 1965 to 116 public hospitals in 2002 along with many private hospitals, health clinics, and rural clinics built throughout the country to provide better health system in Malaysia [20]. As the number of hospitals increased, more facilities were upgraded such as providing more hospital beds. At the same time the number of registered doctors, trained nurses, and midwives have also been increased [20]. In general, the health and medical services in Malaysia have significantly improved since independence contributing to the improvement in infant mortality rates. The government also has put in a lot of effort in terms of the quality of service, the advancement of medicine and medical technologies, the resolution of the issue of unbalanced distributions of medical resources between rural and urban areas, the establishment of collaborations among government and private hospitals or medical institutions, etc. A lot of campaigns and programs have been organized by the local government and the Ministry of Health to educate and increase health consciousness among Malaysians.

In conclusion, as discussed before, even though the space–time mixture model has some advantages in estimating the disease risks and provides a better and clearer picture of categorization, this approach still has a weakness in which the relative risk for different districts could possibly be correlated, i.e., dependent on geographical proximity. An example of the model that can be used which includes the neighboring factor among the areas is the parametric conditional autoregressive model [3].

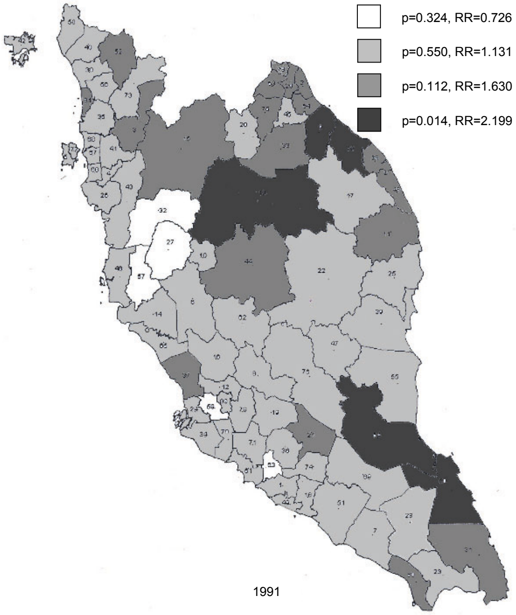


Fig. 8.1. Infant mortality map based on space–time mixture model for the year 1991

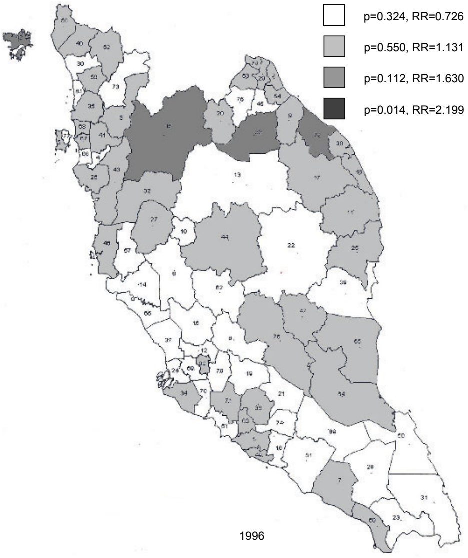


Fig. 8.2. Infant mortality map based on space–time mixture model for the year 1996

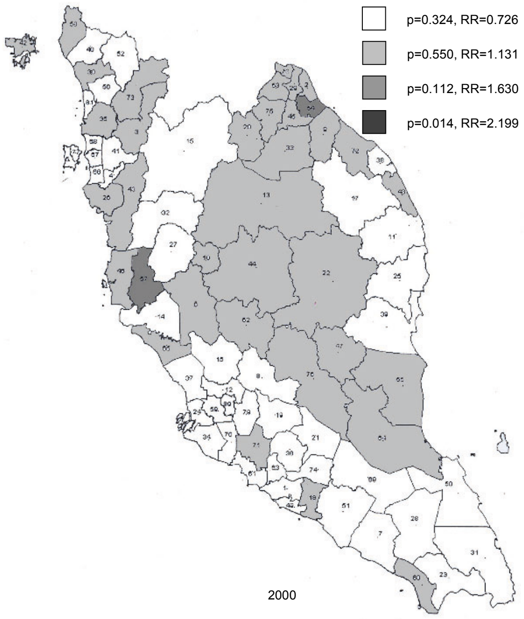


Fig. 8.3. Infant mortality map based on space–time mixture model for the year 2000

References

1. Adebayo SB, Fahrmeir L, Klasen S (2004) Analyzing infant mortality with geoadditive categorical regression model: a case study for Nigeria. Economics and Human Biology 2:229–244
2. Biggeri A, Dreassi E, Lagazio C, Bohning D (2003) A transitional non-parametric maximum pseudo-likelihood estimator for disease mapping. Computational Statistics & Data Analysis 41:617–629
3. Clayton D, Kaldor J (1987) Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. Biometrics 43:671–681
4. Estimates of Malaysia Federal Revenue and Expenditure (1970–2000) Ministry of Finance Malaysia
5. Everitt BS, Hand DJ (1981) Finite mixture distributions. Chapman and Hall, New York
6. Everitt BS, Hand DJ (1993)
7. Langford IH, Leyland AH, Rasbash J, Goldstein H (1999) Multilevel modeling of the geographical distributions of diseases. Applied Statistics 48:253–268

8. Lawson AB, Browne WJ, Rodeiro CLV (2003) Disease mapping with WinBUGS and MlwiN. Wiley, New York
9. Lawson AB, Williams FLR (2001) An introductory guide to disease mapping. Wiley, New York
10. Mantel N, Stark CR (1968) Computation of indirect-adjusted rates in the presence of confounding. Biometrics 24:997–1005
11. Marshall RJ (1991) Mapping disease and mortality rates using empirical Bayes estimators. Applied Statistics 40:283–294
12. Meza JL (2003) Empirical Bayes estimation smoothing of relative risks in disease mapping. Journal of Statistical Planning and Inference 112:43–62
13. Mohamed WN, Diamond I, Smith WFP (1998) The determinants of infant mortality in Malaysia: a graphical chain modeling approach. Journal of the Royal Statistical Society A 161:349–366
14. Pollard AH, Yusuff F, Pollard GN (1981) Demographic techniques. Pergamon Press, Sydney
15. Rattanasiri S, Bohning D, Rojanavipart P, Athipanyakom S (2004) A mixture model application in disease mapping of malaria. Southeast Asian Journal Trop Med Public Health 35:38–47
16. Robbins H (1964) The empirical Bayes approach to statistical decision problems. Annals of Mathematical Statistics 35:1–20
17. Rutstein SO (2005) Effects of preceding birth intervals and neonatal, infant and under-five years mortality and nutritional status in developing countries: evidence from the demographic and health survey. International Journal of Gynecology and Obstetrics 89:7–24
18. Schlattmann P, Bohning D (1993) Mixture models and disease mapping. Statistics in Medicine 12:1943–1950
19. Schlattmann P, Dietz E, Bohning D (1996) Covariate adjusted mixture models and disease mapping wih the program Dismapwin. Statistics in Medicine 15:919–929
20. Social Statistics Bulletin Malaysia (1965–2000) Department of Statistics Malaysia
21. Turrell G, Mengersen K (2000) Socioeconomic status and infant mortality in Australia: a national study of small urban areas, 1985–1989. Social Science & Medicine 50:1209–1225

Chapter 9

A novel hybrid high-dimensional model representation (HDMR) based on the combination of plain and logarithmic high-dimensional model representations

B. Tunga[1], M. Demiralp[2]

[1]Informatics Institute, Istanbul Technical University, 34469 Besiktas, Istanbul, Turkey, burcu@be.itu.edu.tr
[2]Informatics Institute, Istanbul Technical University, 34469 Maslak, Is tanbul, Turkey, demiralp@be.itu.edu.tr

Abstract. This chapter focuses on a new version of hybrid high-dimensional model representation for multivariate functions. High-dimensional model representation (HDMR) was proposed to approximate the multivariate functions by the functions having less number of independent variables. Toward this end, HDMR disintegrates a multivariate function to components which are, respectively, constant, univariate, bivariate, and so on in an ascending order of multivariance. HDMR method is a scheme truncating the representation at a prescribed multivariance. If the given multivariate function is purely additive then HDMR method spontaneously truncates at univariance, otherwise the highly multivariant terms are required. On the other hand, if the given function is dominantly multiplicative then the logarithmic HDMR method which truncates the scheme at a prescribed multivariance of the HDMR of the logarithm of the given function is taken into consideration. In most cases the given multivariate function has both additive and multiplicative natures. If so then a new method is needed. Hybrid high-dimensional model representation method is used for these types of problems. This new representation method joins both plain high-dimensional model representation and

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_9,

logarithmic high-dimensional model representation components via an hybridity parameter. This work focuses on the construction and certain details of this novel method.

Keywords. Multivariate approximation, High-dimensional model representation, Logarithmic HDMR

9.1 Introduction

High-dimensional model representation (HDMR) was first designed by Sobol in 1993 [1]. It is based on the divide-and-conquer philosophy such that the original function is additively represented by a constant term followed by univariate terms, bivariate terms, and so on. So, an N-dimensional multivariate function under consideration can be represented by a constant term, N number of univariate terms, $N(N-1)/2$ bivariate terms, $N(N-1)(N-2)/6$ number of trivariate terms, and so on. Hence, the total number of HDMR components for a given N-variate function is 2^N. Although this number is finite it may climb to very high number as N increases. For example, it contains 2^{100}, approximately 1 million, additive components to have an exact representation for the case of hundred independent variables. This urges us to truncate HDMR at rather small multivariances as long as the truncation has a good representation quality. The general tendency is to truncate, at most, bivariance.

The most important advantage of HDMR method is to deal with less variate functions instead of highly multivariate functions as we have mentioned above. In spite of today's advanced computer technology, the direct valuation of multivariate functions in computers is still fairly difficult especially when the function's dimensionality increases to high values due to the physical limitations on memory and processors. This reality stimulates mathematicians to develop certain methods based on divide-and-conquer philosophy. One of the most recently developed methods in this direction is called high-dimensional model representation (HDMR). HDMR and some other related algorithms were developed in a more comprehensive form by Rabitz and his group [2–5] after Sobol's revolutionary work. Sobol’s suggestion was generalized by Rabitz group such that the integration limits are assumed to be any two real numbers and a weight function which is product of univariate factors, each of which depends on a different independent variable inserted to the integrand as a multiplicative factor. Later, product-type weight function is generalized beyond the Rabitz group's case by using a nonproduct type of weight function under another auxiliary product-type weight function by Demiralp's group. Demiralp and his group developed some other related HDMR methods at the same time period.

Demiralp's group tried to extend HDMR to more general cases to increase its power and efficiency. Amongst the products of these efforts we can mention hybrid HDMR (HHDMR) [8, 9] which combines HDMR and factorized HDMR [6, 7] via a flexible combination parameter. This type of HDMR method works well when the original function has an intermediate nature which corresponds to neither an exactly additive nor an exactly multiplicative nature. In this work, a new HHDMR expansion including logarithmic HDMR instead of factorized HDMR again under a hybridity parameter is proposed. The main idea here is to get rid of the main disadvantage of the FHDMR structure, which is about the definition of the multiplicativity measurers. The structure developed in logarithmic HDMR method allows us to define new truncation quality measurers which are monotonously increasing from 0 to 1 in ascending multivariance. This feature furnishes us with better understanding of the behaviors and qualities of the HHDMR approximants.

The rest of this chapter is organized as follows. The second section is about HDMR to recall the construction details of the method. The third section presents the basic idea underlying logarithmic HDMR (LHDMR). The fourth section presents the core of this chapter, "A new hybrid approach in high-dimensional model representations (HHDMR)." The fifth section contains simple illustrative applications for this new hybrid approach in HDMR and the sixth section finalizes the chapter with concluding remarks.

9.2 HDMR

The high-dimensional model representation [1–10] of a multivariate function $f(x_1,\ldots,x_N)$ is given as

$$\begin{aligned} f(x_1,\ldots,x_N) = f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N} f_{i_1,i_2}(x_{i_1},x_{i_2}) \\ + \cdots + f_{1\ldots N}(x_1,\ldots,x_N) \end{aligned} \tag{9.1}$$

where N stands for the number of the independent variables and the right-hand-side components are orthogonal in an Hilbert space over the hyperprism defined by the intervals $a_i \le x_i \le b_i$ (where $1 \le i \le N$ and a_i,

b_i are assumed to be given). The inner product in Hilbert space is defined as follows for two arbitrary square integrable multivariate functions $g(x_1,\ldots,x_N)$ and $h(x_1,\ldots,x_N)$:

$$(g,h)=\int_{a_1}^{b_1}dx_1\cdots\int_{a_N}^{b_N}dx_N W(x_1,\ldots,x_N)g(x_1,\ldots,x_N)h(x_1,\ldots,x_N) \tag{9.2}$$

where $W(x_1,\ldots,x_N)$ stands for a product-type function and it can be given as follows:

$$W(x_1,\ldots,x_N)\equiv\prod_{i=1}^{N}W_i(x_i) \tag{9.3}$$

Here we assume that $W_i(x_i)$ ($1\le i\le N$), the components of $W(x_1,\ldots,x_N)$, are given and the integral of these components between a_i and b_i is equal to 1. These weight factors must be chosen to fulfill the requirement for being true weight functions (they should be either always positive everywhere or always negative everywhere except at certain finite number of points where they may vanish). Otherwise the monotonic increasing nature in truncation quality measurers for ascending multivariance may disappear.

The HDMR components in the right-hand side of Eq. (9.1) can be determined uniquely by imposing mutual orthogonality amongst these components. This feature allows us to determine constant term f_0 by using the following projection operator:

$$P_0 g(x_1,\ldots,x_N)\equiv\int_{a_1}^{b_1}dx_1\cdots\int_{a_N}^{b_N}dx_N W(x_1,\ldots,x_N)g(x_1,\ldots,x_N) \tag{9.4}$$

The orthogonality of all higher than zero-order multivariate components to f_0 implies that the integrals of those components over one of their independent variables over the related interval under the corresponding univariate weight function vanish (vanishing property proposed by Sobol). Now if we apply the projection operator P_0 on both sides of Eq. (9.1) and then utilize the vanishing properties of the higher than zero variate terms and the normalized nature of the weight function factors, we can write

$$f_0=P_0 f(x_1,\ldots,x_N) \tag{9.5}$$

To determine the univariate terms, $f_i(x_i)$ s, by using the orthogonality feature we have to define projection operators P_i ($1\le i\le N$). They are equivalent to P_0's new forms obtained after removing the integration over x_i and discarding the univariate weight function factor $W_i(x_i)$. If we apply the action of P_i on both sides of Eq. (9.1) then the employment of the vanishing properties of all HDMR terms except the constant one and the normalization in univariate weight factors enables us to write

$$f_i(x_i) = P_i f(x_1,\ldots,x_N) - f_0, \qquad 1 \le i \le N \tag{9.6}$$

As can be easily seen the determination of bivariate and higher multivariate HDMR components can be realized by defining other projection operators $P_{i_1 \cdots i_k}$ ($1 \le i \le N$). We do not intend to explicitly give them here. By using these operators the higher order HDMR terms can be obtained in a similar manner.

It is not hard to see from the HDMR equation given in Eq. (9.1) that the schemes based on HDMR truncations are finite-step methods. However, working with all HDMR components becomes a nightmare when the dimensionality increases to high values as we have mentioned above. This is because of the exponential growth, 2^N, with respect to N in the number of HDMR terms. To avoid this problem HDMR equation Eq. (9.1) can be truncated at some level of multivariance (preference is at most to keep bivariate terms). These truncations are called HDMR approximants and are given below:

$$\begin{aligned}
&s_0(x_1,\ldots,x_N) = f_0 \\
&s_1(x_1,\ldots,x_N) = s_0(x_1,\ldots,x_N) + \sum_{i=1}^{N} f_i(x_i) \\
&\vdots \\
&s_k(x_1,\ldots,x_N) = s_{k-1}(x_1,\ldots,x_N) + \sum_{\substack{i_1 \cdots i_k = 1 \\ i_1 < \cdots < i_k}}^{N} f_{i_1 \cdots i_k}(x_{i_1},\ldots,x_{i_k}) \\
&1 \le k \le N
\end{aligned} \tag{9.7}$$

The next step is to measure the quality of these approximants for the characterization of the original function within a desired numerical precision. Entities which are called “additivity measurers” are defined for this purpose:

$$\begin{aligned}
\sigma_0 &= \frac{1}{\|f\|^2}\|f_0\|^2 \\
\sigma_1 &= \frac{1}{\|f\|^2}\sum_{i=1}^{N}\|f_i\|^2 + \sigma_0 \\
&\vdots \\
\sigma_N &= \frac{1}{\|f\|^2}\|f_{12\cdots N}\|^2 + \sigma_{N-1}
\end{aligned} \tag{9.8}$$

Here, σ_0 is called "constancy measurer" and it defines the contribution percentage of the constant term to the HDMR expansion's norm square. σ_1 is called "first-order additivity measurer" and it defines the contribution percentage of the constant term and univariate terms to the HDMR expansion's norm square. As a generalization σ_k called "k_{th} order additivity measurer" defines the contribution percentage of all the terms from constant term to k_{th} -order term inclusive of the HDMR expansion's norm.

As we mentioned earlier, it is very hard to construct truncation quality measurers that monotonically increase as the multivariance ascends in the case of FHDMR although it is possible to truncate the finite term product at certain level of multivariance. As a matter of fact such measurers could not be constructed. To avoid this difficulty the logarithmic HDMR (LHDMR) has been developed.

9.3 Logarithmic HDMR

Logarithmic high-dimensional model representation method is based on the idea of expanding the natural logarithm of a nonnegative multivariate function to HDMR instead of the function's itself. LHDMR formula which defines a product-type representation for a given multivariate function can be expressed as follows:

$$\begin{aligned}
\ln\left[f(x_1,\ldots,x_N)-\Phi(x_1,\ldots,x_N)\right] = \varphi_0 + \sum_{i_1=1}^{N}\varphi_{i_1}(x_{i_1}) + \\
\sum_{\substack{i_1,i_2=1\\ i_1<i_2}}^{N}\varphi_{i_1,i_2}(x_{i_1},x_{i_2}) + \cdots
\end{aligned} \tag{9.9}$$

where $\Phi\left(x_1,\ldots,x_N\right)$ is a minorant function to the given function, $f\left(x_1,\ldots,x_N\right)$, to produce a nonnegative or preferably positive core function for the logarithm. We call this entity "reference function" since it takes somehow the role of the origin in the space of the functions. The right-hand-side components of Eq. (9.9) are mutually orthogonal and can be determined by tracing the basic rule of the HDMR method.

If Eq. (9.9) is reorganized the following representation formula for LHDMR is obtained:

$$f(x_1,\ldots,x_N)=\Phi(x_1,\ldots,x_N)+e^{\varphi_0}\left[\prod_{i_1=1}^{N}e^{\varphi_{i_1}(x_{i_1})}\right]\times \left[\prod_{\substack{i_1,i_2=1\\ i_1<i_2}}^{N}e^{\varphi_{i_1,i_2}(x_{i_1},x_{i_2})}\right]\times\cdots \tag{9.10}$$

The explicit expressions of LHDMR approximants can be written as follows when the minorant function is assumed to be vanishing for simplicity (otherwise they will be more complicated although a recursive structure can be constructed):

$$\begin{aligned}
&\lambda_0(x_1,\ldots,x_N)=e^{\varphi_0}\\
&\lambda_1(x_1,\ldots,x_N)=\lambda_0(x_1,\ldots,x_N)\prod_{i_1=1}^{N}e^{\varphi_{i_1}(x_{i_1})}\\
&\lambda_k(x_1,\ldots,x_N)=\lambda_{k-1}(x_1,\ldots,x_N)\prod_{\substack{i_1\cdots i_k=1\\ i_1<\cdots<i_k}}^{N}e^{\varphi_{i_1\cdots i_k}(x_{i_1},\ldots,x_{i_k})}\\
&1\le k\le N
\end{aligned} \tag{9.11}$$

LHDMR method allows us to define the following truncation quality measurers:

$$
\begin{aligned}
v_0 &= \frac{\|\varphi_0\|^2}{\|\ln(f-\Phi)\|^2} \\
v_1 &= \frac{\|\varphi_0\|^2 + \sum_{i_1=1}^{N}\|\varphi_{i_1}\|^2}{\|\ln(f-\Phi)\|^2} \\
v_2 &= \frac{\|\varphi_0\|^2 + \sum_{i_1=1}^{N}\|\varphi_{i_1}\|^2 + \sum_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N}\|\varphi_{i_1,i_2}\|^2}{\|\ln(f-\Phi)\|^2} \\
&\vdots
\end{aligned}
\tag{9.12}
$$

The following inequality holds for these measurers:

$$0 \le v_0 \le v_1 \le \cdots \le v_N \le 1 \tag{9.13}$$

Until this point we have presented some preliminary information about HDMR methods appearing in the new version of the HHDMR. Now we have been sufficiently equipped to present the novel version of HHDMR.

9.4 New version of hybrid HDMR

The multivariate functions which are neither dominantly additive nor dominantly multiplicative push us to develop a hybrid algorithm. Previously developed hybrid HDMR joins plain HDMR and FHDMR methods under a hybridity parameter. In this work, logarithmic HDMR takes the role of factorized HDMR. Hence, hybrid HDMR method to be presented here has a new expansion including both HDMR and LHDMR expansions via a hybridity parameter now:

$$
\begin{aligned}
f(x_1,\ldots,x_N) &= \alpha\left(f_0 + \sum_{i_1=1}^{N} f_{i_1}(x_{i_1}) + \cdots\right) + \\
&(1-\alpha)\left(\Phi(x_1,\ldots,x_N) + e^{\varphi_0}\left[\prod_{i_1=1}^{N} e^{\varphi_{i_1}(x_{i_1})}\right]\left[\prod_{\substack{i_1,i_2=1 \\ i_1<i_2}}^{N} e^{\varphi_{i_1,i_2}(x_{i_1},x_{i_2})}\right]\times\cdots\right)
\end{aligned}
\tag{9.14}
$$

where α is the hybridity parameter. We can define the following approximants through this definition:

$$\begin{aligned} f(x_1,\ldots,x_N) &\approx h_{jk}(x_1,\ldots,x_N;\alpha) \\ &= \alpha s_j(x_1,\ldots,x_N) + (1-\alpha)\lambda_k(x_1,\ldots,x_N), \\ 1 \le j,k \le N \end{aligned} \tag{9.15}$$

This approximant is called the (j,k)th -order hybrid HDMR approximant. An $(N+1)\times(N+1)$ table like Pade Ratios can be constructed and then used for approximating the original function:

$$\begin{matrix} h_{00} & h_{01} & \cdots & h_{0N} \\ h_{10} & h_{11} & \cdots & h_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ h_{N0} & h_{N1} & \cdots & h_{NN} \end{matrix} \tag{9.16}$$

The approximating capability of each HHDMR approximant can be defined as follows:

$$q_{jk} = \frac{\left\| f - h_{jk} \right\|^2}{\left\| f \right\|^2} \tag{9.17}$$

which is somehow error bound. The best approximating capability is of course 0 for these approximants.

9.5 Implementations

To illustrate how HHDMR works we can choose a multivariate function whose additivity and multiplicativity can be controlled by a single integer parameter as follows:

$$f(x_1,\ldots,x_N) = (x_1 + \cdots + x_N)^m \tag{9.18}$$

where m is an integer varying between 0 and N inclusive. The function above starts from the constant value when $m=1$ and its multivariance increases through univariance, bivariance, and so on as m increases one by one as long as the HDMR's geometry is taken as a hyperprism whose one corner is located in the origin of the Cartesian space spanned by the independent variables. Hence in the case of $m=1$ it is purely additive and its multiplicativity increases as m grows although pure multiplicativity is never achieved.

If we use Eq (9.15) in Eq. (9.17) then we obtain the following formula:

$$q_{jk} = \frac{\left\| f - \lambda_k \right\|^2}{\left\| f \right\|^2} + 2\alpha \frac{(f - \lambda_k, \lambda_k - s_j)}{\left\| f \right\|^2} + \alpha^2 \frac{\left\| \lambda_k - s_j \right\|^2}{\left\| f \right\|^2}, \qquad (9.19)$$

$$0 \le j,k \le N$$

The optimization of this entity with respect to α gives the following unique result for the optimum value of α:

$$\alpha_{j,k}^{(opt)} = \frac{(f - \lambda_k, s_j - \lambda_k)}{\left\| \lambda_k - s_j \right\|^2}, \qquad 0 \le j,k \le N \qquad (9.20)$$

As can be noticed easily this value turns out to be 1 when $j = 0$, $k = 0$, and f is a constant. However, it differs from one in the other cases. Although there is no warranty that it will stay between 0 and 1 one can investigate the situation and try to find which kind of functions gives optimized α values in $[0, 1]$. Our observations show that our test function behaves in this manner. However, this may not be true for some other multivariate functions which are neither purely additive nor purely multiplicative. We do not intend to further details of this issue here. Also we do not report the details of our observations here due to space constraint.

9.6 Conclusion

This work is devoted to the construction of a more efficient version of HHDMR. This has been necessary to replace FHDMR, which has no truncation quality measurers increasing parallel to the increase in multivariance, with LHDMR which has such kind of measurers. LHDMR is based on the expansion of a multivariate function's logarithm to plain HDMR and is based on the fact that logarithm converts multiplicativity to additivity and this is why it is used here.

The implementation results encourage us to use HHDMR in the approximation of the functions which are not dominantly additive or multiplicative.

Acknowledgment. The first author is grateful to the State Planning Organization (DPT) of Turkey for its support and the second author thanks Turkish Academy of Sciences for its support and motivation.

References

1. Sobol IM (1993) Sensitivity estimates for nonlinear mathematical models. Mathematical Modelling and Computational Experiments (MMCE) 1:407
2. Shorter JA, Ip PC, Rabitz H (1999) An efficient chemical kinetics solver using high dimensional model representation. Journal of Physical Chemistry. A 103:7192–7198
3. Alış Ö, Rabitz H (2001) Efficient implementation of high dimensional model representations. Journal of Mathematical Chemistry. 29:127–142
4. Li G, Wang SW, Rabitz H, Wang S, Jaffe P (2002) Global uncertainty assessments by high dimensional model representation (HDMR). Chemical Engineering Science 57:4445–4460
5. Li G, Artamonov M, Rabitz H, Wang S, Georgopoulos PG, Demiralp M (2003) High dimensional model representations generated from low order terms -lp-RS-HDMR. Journal of Computational Chemistry 24:647–656
6. Tunga MA, Demiralp M (2004) A factorized high dimensional model representation on the partitioned random discrete data. Applied Numerical Analysis and Computational Mathamatics. 1:231–241
7. Tunga MA, Demiralp M (2005) A factorized high dimensional model representation on the nodes of a finite hyperprismatic regular grid. Applied Mathematics and Computation 164:865–883
8. Tunga MA, Demiralp M (2006) Hybrid high dimensional model representation (HHDMR) on the partitioned data. Journal of Computational and Applied Mathematics 185:107–132
9. Tunga B, Demiralp M (2003) Hybrid high dimensional model representation approximants and their utilization in applications. Mathematical Research 9:438–446
10. Demiralp M (2003) High dimensional model representation and its varieties. Mathematical Research 9:146–159

Chapter 10

A decision support system to evaluate the competitiveness of nations

Ş.Önsel,[1] F.Ülengin,[1] G.Ulusoy,[2] Ö.Kabak,[3] Y.İ.Topcu,[3] E.Aktaş [3]

[1] Department of Industrial Engineering, Dogus University, Acibadem, Kadikoy, 34722, Istanbul, Turkey, {sonsel, fulengin}@dogus.edu.tr
[2] Faculty of Engineering and Natural Sciences, Sabanci University, Orhanli, Tuzla 81474 Istanbul, Turkey, gunduz@sabanciuniv.edu
[3] Department of Industrial Engineering, Istanbul Technical University, Macka, 34357 Istanbul, Turkey, {kabak, topcuil, aktasem}@itu.edu.tr

Abstract. The aim of this chapter is to explore methodological transparency as a viable solution to problems created by existing aggregated indices as well as to conduct a detailed analysis on the ongoing performance of nations' competitiveness. For this purpose, a methodology composed of three steps is used. To start with, a combined clustering analysis methodology is used to assign countries to appropriate clusters. Unlike the current methods that use a single criterion, the proposed methodology uses 135 criteria for a proper classification of the countries. Relationships between the criteria and classification of the countries are determined using artificial neural networks (ANNs). ANN provides an objective method for determining the criteria weights, which are, for the most part, subjectively specified in existing methods. Finally, the countries are ranked based on weights generated in the previous step. As a final analysis, the dynamic change of the rank of the countries over years has also been investigated.

Keywords. Ranking, Competitiveness, Artificial neural network, Cluster analysis

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_10,

10.1 Introduction

A nation's competitiveness can be viewed as its position in the international marketplace compared to other nations of similar economic development. The capability of firms to survive and to have a competitive advantage in global markets depends, among other things, on the efficiency of their nation's public institutions, excellence of the educational, health, and communication infrastructures, as well as the nation's political and economical stability. On the other hand, an outstanding macroeconomic environment alone cannot guarantee a high level of national competitive position unless firms create valuable goods and services with a commensurately high level of productivity at the micro-level. Therefore, the micro- and macroeconomic characteristics of an economy jointly determine its level of productivity and competitiveness.

Although many view competitiveness as a synonym for productivity [1], these two related terms are, in fact, different. Productivity refers to the internal capability of an organization while competitiveness refers to the relative position of an organization vis-à-vis its competitors. Each year, some organizations, such as the World Economic Forum (WEF) [2] and the Institute for Management Development (IMD) [3], publish rankings of national competitiveness among countries. These rankings serve as benchmarks for national policy makers and interested parties in judging the relative success of their countries in achieving competitiveness as represented by well-known and accepted indices. However, for the last quarter-century, the WEF has led in evaluation of the competitiveness of nations through its publication, The Global Competitiveness Report [2].

With the 2006–2007 report [2], WEF decided to use the global competitiveness Index (GCI) [4], as the main competitiveness indicator. The GCI, albeit simple in structure, provides a holistic overview of factors that are critical to driving productivity and competitiveness and groups them into nine pillars that are different from the 2004–2005 report [5] where 12 pillars are assumed. Combining some of the pillars and separating a pillar result in such a decrease.

The nine pillars are measured using both hard data from public sources (such as inflation, Internet penetration, and school enrolment rates) and data from the World Economic Forum's Executive Opinion Survey, which is conducted annually among top executives in all of the countries assessed. The survey provides crucial data on a number of qualitative issues (e.g., corruption, confidence in the public sector, quality of schools) for which no hard data exist.

In addition to the change in the number of pillars, there are also changes on the variable configuration and structure of the pillars. In the recent report the total number of variables is decreased from 177 to 137.

Although there are changes in the pillars and variables in the 2006–2007 report, the basic approach to the evaluation procedure remains unchanged. The pillars are still used as sub-index for three main dimensions of competitiveness: basic requirements (the first four pillars), efficiency enhancers (fifth to seventh pillars), and innovation and sophistication factors (last two pillars).

An important characteristic of the GCI is that it explicitly takes into account the fact that the countries around the world are at different levels of economic development. What is important for improving the competitiveness of a country at a particular stage of development will not necessarily be the same for a country in another stage. Thus GCI separates countries into three specific stages: factor-driven, efficiency-driven, and innovation-driven. Therefore, in the calculation of the final GDI, the weights of the three dimensions are determined according to the stage that country belongs to [2]. Unfortunately, this classification tends to be rather subjective or is based solely on per capita income. Subjectivity is also present when creating the threshold used to separate one stage from another. Some degree of objectivity is possible, however, if countries are clustered as a function of their similarities on selected criteria. By doing so, important factors underlying the competitiveness position of each stage, and of particular countries at various stages, can be revealed. It will thus be easier to understand the internal dynamics of each stage and to provide useful and objective guidelines to countries as they attempt to improve their positions with respect to those located at higher stages.

Section 10.2 of this chapter introduces our proposed methodology to cluster countries into stages and to generate criteria weights that are critical at each stage of the procedure. In Sect. 10.3, a composite index is calculated using the calculated weights. The results are then compared to those of the GCI of the WEF to determine whether the weights adopted by the WEF incorrectly penalize some countries and/or reward others. Besides, in this section the change of the nations' competitiveness rank is also analyzed. This chapter closes with conclusions and suggestions for further improvements of the proposed methodology.

10.2 Proposed methodology

The aim of this research is, first, to provide an *objective clustering of countries* according to their values/scores on selected criteria and, second, to propose an *objective weighting procedure to calculate an aggregated index.* For these purposes, a three-step methodology is proposed. Finally the results are compared with our previous study's [6] findings to track the changes in the competitiveness of the countries between 2005 and 2007.

The proposed methodology considers 135 criteria in the clustering process. The criteria are the hard data and survey data used in the WEF report [2]. Two of the criteria have been removed from consideration due to lack of available data.

In particular, a hierarchical cluster analysis is used to determine the "best" number of clusters; this number is then used as a parameter to determine the appropriate clusters of countries using self-organizing maps [7]. Next, relationships between the criteria and the classification of countries are determined in an objective manner using artificial neural networks (ANN). Importantly, existing methodologies generally assess criteria weights/importances subjectively. Finally, in the third step of our procedure, countries are rank-ordered based on the ANN-generated weights and the dynamic change in the rank of countries is analyzed. The proposed methodology can also be used to identify those attributes a country should focus on in seeking to improve its position relative to other countries, i.e., to transition from its current cluster to a better one.

10.2.1 Classification of countries

In the first part of this research, countries are grouped based on their similarity of characteristics. Cluster analysis is used for this purpose.

10.2.1.1 Cluster analysis

Cluster analysis involves grouping similar objects into mutually exclusive subsets referred to as clusters [7]. The cluster definition problem is NP-complete, so a computationally efficient, exact solution method, to the best of the authors' knowledge, does not exist. However, a number of heuristic methods have been proposed for this purpose, including agglomerative techniques [7]. All hierarchical agglomerative heuristics begin with **n** clusters, where **n** is the number of observations. Then, the two most similar clusters are combined to form **n−1** clusters. On the next iteration, **n−2** clusters are formed with the same logic, and this process continues until one cluster remains. Only the rules used to merge clusters differ across the various heuristics.

In order to improve the accuracy of, and reduce any subjectivity in, the cluster analysis, we employ a self-organizing map (SOM) neural network, as suggested by Mangiameli et al [8]. The SOM is thus not taken as an alternative, but rather as a complementary analysis that follows hierarchical clustering. The focus is on improved accuracy in the assignment of observations to appropriate clusters, given that the number of clusters in the data is known. The SOM's network learns to detect groups of similar input vectors in such a way that neurons physically close in the neuron layer respond to a similar input vector [9].

10.2.1.2 Determining the country clusters

The basic drawback of any study based solely on ranking is that the ordinal scale does not reflect the appropriate competitiveness level of a country relative to other countries. The most accurate position of a country within the total configuration can only be determined after the grouping of nations is performed, and similarities to the evaluated country in terms of competitiveness are identified.

In the current study, the Ward hierarchical method, an agglomerative clustering technique, and the Euclidean distance measure were selected as most appropriate based on evaluations using MATLAB [10]. In Ward's method, the distance is the ANOVA sum of squares between two clusters summed over all variables [7]. An analysis of the dendrogram and ANOVA were thus used to test the significance of differences between the cluster means, producing three significant clusters. Dendrogram analysis generates a dendrogram plot of the hierarchical, binary cluster tree. It consists of many U-shaped lines connecting objects in a hierarchical tree. The height of each U represents the distance between the two objects being connected. Each leaf in the dendrogram corresponds to one data point. As can be seen from Fig. 10.1, the countries can be grouped into three different U-shaped clusters according to 2006–2007 data.

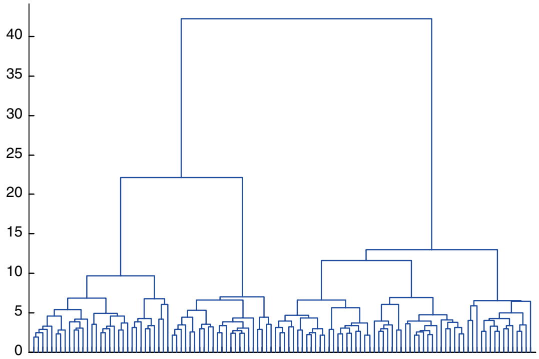


Fig. 10.1. Dendrogram of the country clusters

Next, the appropriate number of clusters generated in the first stage was used to repeat the analysis using SOM and MATLAB software. Since we sought to categorize the countries into three classes, there were three outputs in the ANN's configuration. This generated a 3*1 matrix of the weight vector. The topology function used was "HEXTOP," which means that the neurons were arranged in a hexagonal topology at the Kohonen layer, while the distance function was "MANDIST," i.e., the Manhattan (city block) distance. The training of a self-organizing map using MATLAB involved two steps: ordering phase and tuning phase. In the former, the ordering phase learning rate and neighborhood distance are decreased from the rate and maximum distance between two neurons to the tuning phase learning rate and tuning phase neighborhood distance, respectively. The ordering phase lasts for a given number of steps. At the tuning phase, the learning rate is decreased much more slowly than in the ordering phase, while the neighborhood distance stays constant [11]. In the current study, the ordering phase learning rate, ordering phase steps, and tuning phase learning rate were taken as 0.9, 1000, and 0.02, respectively. The countries contained within the resulting clusters are determined by the end of this first stage of the method. The resulting clusters as well as the related countries that are found by 2006 data are given in Table 10.1.

Table 10.1. Clusters of countries

2006	2004		
	High competitive (HC)	Competitive (CO)	Non-competitive (NC)
HC	Australia, Austria, Belgium, Canada, Denmark, Finland, France, Germany, Iceland, Ireland, Israel, Japan, Luxembourg, Netherlands, New Zealand, Norway, Singapore, Sweden, Switzerland, Taiwan, UK	Chile, Czech Republic, Estonia, India, Korea Rep., Malaysia, Portugal, Slovenia, Spain, Tunisia, United Arab Emirates	
CO		Bahrain, Brazil, Costa Rica, Cyprus, Egypt, Greece, Hungary, Indonesia, Italy, Jordan, Kuwait, Lithuania, Malta, Mauritius, Slovak Rp, S. Africa, Thailand	Colombia, Croatia, El Salvador, Jamaica, Mexico, Panama, Poland, Turkey, Uruguay
NC		Botswana, China, Morocco, Namibia	Algeria, Angola, Argentina, Bangladesh, Bolivia,

		Bosnia and Herz., Bulgaria, Chad, Dominican Rp, Ecuador, Ethiopia, Gambia, Georgia, Guatemala, Honduras, Kenya, Macedonia, Madagascar, Malawi, Mali, Mozambique, Nicaragua, Nigeria, Pakistan, Paraguay, Peru, Philippines, Romania, Russian Fd, Serbia and Montenegro, Sri Lanka, Tanzania, Trinidad and Tobago, Uganda, Ukraine, Venezuela, Vietnam, Zambia, Zimbabwe

Comparison of the results with our previous findings [6] shows that Turkey, our home country, has shown an impressive improvement in the competitive performance moving up from the non-competitive to competitive countries. Colombia, Croatia, El Salvador, Jamaica, Mexico, Panama, Poland, and Uruguay have also shown a similar transition. However, China, Morocco, Namibia, Botswana have moved from the competitive to non-competitive stage. Finally, Chile, Czech Republic, Estonia, India, Korea, Malaysia, Portugal, Slovenia, Spain, Tunisia, United Arab Emirates moved up from the competitive to highly competitive stages. None of the countries previously assigned to highly competitive cluster moved down to a lower stage.

10.2.2 Identification of basic criteria underlying country stages through ANN

At this step of the study, the basic factors underlying the reasons a country belongs to a specific cluster is analyzed using ANN. The feed-forward back propagation algorithm is used for this purpose.

10.2.2.1 Artificial neural networks

ANN techniques have been applied to a variety of problem types and, in many instances, provided superior results to conventional methods [12]. The literature [e.g., 13–15] suggests the potential advantages of ANN versus classical statistical methods. The basic ANN model consists of computational units that emulate the functions of a nucleus in a human brain. The unit receives a weighted sum of all its inputs and computes its own output value by a transformation, or output, function. The output value is then propagated to many other units via connections between units. The learning process of ANN can be thought of as a reward and punishment mechanism [16]. When the system reacts appropriately to an input, the related weights are strengthened. As a result, it becomes possible to generate outputs, which are similar to those of the previously encountered inputs. In contrast, when undesirable outputs are produced, the related weights are reduced. The model will thus *learn* to give a different reaction when similar inputs occur. In this way, the system is "trained" to produce desirable results while "punishing" undesirable ones.

In multilayer networks, all inputs are related to outputs through hidden neurons–i.e., there is no direct relationship among them. As a result, specification of the characteristics of each input neuron and the strength of relation between input X_i and output O_i can be found using the method proposed by Onsel et al. [17]:

$$RS_{ji} = \frac{\left[\sum_{k=0}^{n}(W_{ki} * U_{jk})\right]^2}{\sum_{i=0}^{m}\left[\sum_{k=0}^{n}(W_{ki} * U_{jk})\right]^2} \tag{10.1}$$

In this expression, RS_{ji} represents the strength of relation between input *i* and output j. W_{ki} is the weight between the *j*th output U_{jk} and the *k*th hidden neuron. RS_{ji} is thus the ratio of the strength of relation between the *i*th input and *j*th output to the sum of all such strengths. The absolute value in the denominator is used to avoid positive relations canceling the impact of negative ones.

10.2.2.2 Determining basic criteria weights

Output from the SOM in the previous stage helps generate the clusters of countries. These data are then used as the output of the multilayer feed-forward ANN while the 135 criteria are treated as inputs.

About 68 countries are used for training, 22 countries for validation, and again 22 countries for testing stages. In order to obtain robust results based on different trials, for each hidden neuron number, the ANN is computed 10 times, and the best results obtained from each taken. In this way, an attempt is made to detect different points of weight space corresponding to the network via several experiments. The optimal hidden neuron number is found as 5. The tangent sigmoid function (tansig) is used to show the relation between the input-hidden and the hidden-output layers. The training algorithm is a gradient-descent method with momentum and an adaptive learning ratio (traingdx). The validation vectors are used to stop training early if further training on the primary vectors will hurt generalization to the validation vectors [11]. Test vector performance can be used to measure how well the network generalizes beyond primary and validation vectors. The mean square error, selected as the performance measurement, was found to be 0.00017. The importance of the inputs (criteria), playing the dominant role in allocation of countries to the three clusters, was obtained using the modified Onsel et al. [17] formula. The most important five criteria in each cluster are as follows:

High-Competitive Countries: inflation, local equity market access, reliance on professional management, personal computers, local supplier quantity

Competitive Countries: judicial independence, pervasiveness of illegal donations to political parties, medium-term business impact of tuberculosis, medium-term business impact of malaria, informal sector.

Non-competitive Countries: Local equity market access, favoritism in decisions of government officials, degree of customer orientation, local supplier quantity, pay and productivity

10.3 Ranking countries based on the proposed weighted criteria index

At the third step of this research, the weights of 135 criteria for each cluster calculated in the previous step are used to rank the countries. For this purpose, initially, the weights are normalized. The score obtained by each country from each of the criteria is then multiplied by the normalized weight of that criterion. The 112 countries are subsequently ranked according to these weighted index values.

The top 10 ranked countries are (in the order of rank) Switzerland, Finland, Denmark, Germany, Sweden, Singapore, Japan, Netherlands, Hong Kong SAR, UK, Austria, and United States.

The rankings found with the proposed approach are compared with that obtained with 2004–2005 data. This dynamic comparison will also show which countries have dealt with the key determinants of competitiveness at their level of development such as macroeconomic stability or education and health. It is also possible to underline the additional factors over which the countries should focus in order to switch to higher clusters of development.

In Fig. 10.2, the countries that have an improvement or decline by 10 points with respect to 2005 ranking can be seen. Accordingly, India had the most dramatic improvement moving from 49th to 29th in ranking. Poland (+17) and Guatemala (+20) constitute the other countries having the most important improvement in ranking. Besides, our home country, Turkey, moving from 49th to 60th is among the most improved seven countries.

On the other hand Namibia is a country which shows a dramatic decrease with respect to previous analysis ranking (from 78th to 39th). It is interesting to note that China (−17) and Brazil (−16) which are accepted as emerging countries as well as Bulgaria which has recently accessed EU countries are among countries with worst competitiveness performance change with respect to 2004 ranking.

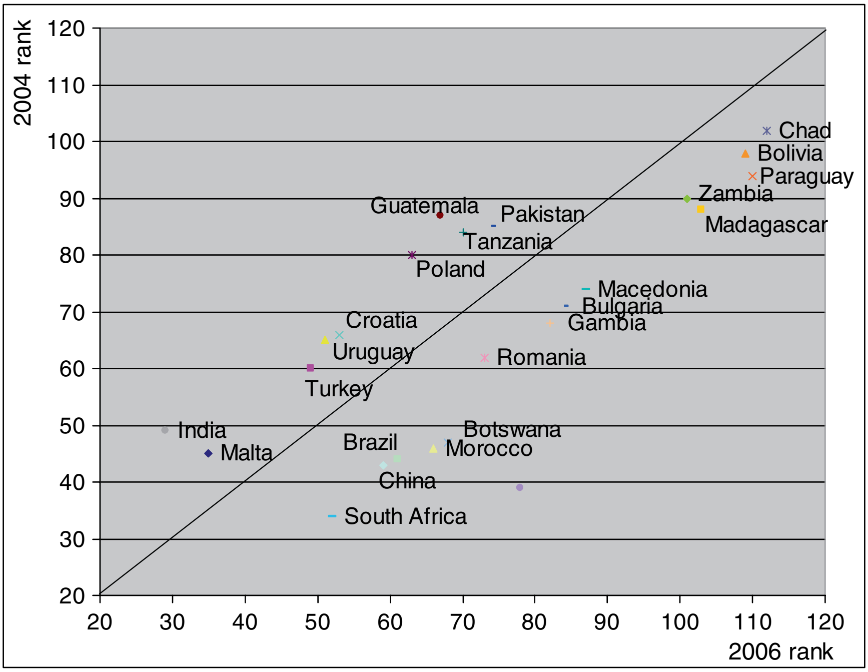


Fig. 10.2. Countries that show important changes in ranking (2004–2006)

The rankings found with the proposed approach are compared with that obtained using the WEF's GCI. According to WEF, the countries having a GDP below a threshold level are accepted as Stage 1 countries and their key factors are assumed to be the basic requirement factors.

However, this is a non-compensatory approach and there may be some countries showing very good performance in terms of basic requirements while still having a low level of GDP. Therefore, it may be unfair to assign a country to a stage based solely on its GDP level and it may be more accurate to use a compensatory approach for this purpose.

A country may be unfairly rewarded due to its high GDP level, although it has poor performance even in terms of its basic requirement factors. For example, the United States does not score well (27th) in terms of basic requirements. However, it is the world's leader in both efficiency enhancers and innovation and sophistication factors. This is mainly due to the fact that the United States is in the third stage of development (the innovation stage) and the weight of the basic requirements is relatively minor. Therefore the high values that it receives from the other two sub-indexices put this country in the leading position.

It is important to emphasize that the subjectivity of the WEF clustering, as well as of the weighting process, sometimes results in contradictory results with respect to the WEF's index. In particular, important discrepancies may occur between the stage to which a country is assigned and the rank that it receives based on the GCI. When a country is assigned to a stage, logically, it is not expected to be ranked lower than the countries in worst stages nor higher than the ones in better stages.

WEF results show some contradictions according to this perspective while our methodology provides a complete parallelism between the stage to which a country is assigned and its place with respect to the global ranking. For example, Taiwan is assigned by WEF to transition stage between Stage 2 (efficiency-driven economies) and Stage 3 (innovation-driven economies), while it is found to be the 13th country according to the GCI. However, our method assigned Taiwan to high-competitive country cluster and found appropriately that its rank is 18th. Tunisia is assigned by WEF as in transition stage between 1 and 2 while its rank is 30 which is too high for such a low-stage cluster. However, our method correctly classified it as high-competitive country and assigned it to 30th rank.

Contrarily, Italy, Kuwait, Cyprus, and Greece were assigned by WEF as Stage 3 (high-competitive) countries, but their ranks are between 42nd and 47th. However, our method correctly assigned them to competitive rather than high-competitive stage as expected.

10.4 Conclusion and further suggestions

Despite attempts to provide objectivity in the development of indicators for the analysis of the competitiveness of countries, there are obviously subjective judgments about how data sets are aggregated and what weighting is applied. Generally, either equal weighting is applied to calculate the final index or subjective weights are specified. The same problem also occurs in the subjective assignment of countries into different clusters. For example the WEF assigns countries to different stages of development mainly on the basis of their GDP level and the application of different subjective weights for each stage. These subjectivities may create a bias, as selecting specific data simultaneously overestimates the level of competitiveness of some countries, making them look unrealistically good, while underestimating that of others.

The aim of this chapter is to explore whether methodological transparency can be an adequate solution to the above-given problems posed by the current aggregated indices. For this purpose, a methodology is proposed to objectively group countries into clusters as well as to specify the weight of the criteria that play the dominant role in each cluster. A new composite index that uses calculated weights has been created. By doing so, the criticism that it is simply an attempt to make some countries more competitive than they actually are can be avoided. What is more, by focusing on the criteria necessary to move a country into a higher cluster, the index can be used by both policy makers and executives responsible for making their countries more competitive.

Moreover, the dynamic structure of the changes in the rankings of the countries' competitiveness level is also analyzed in detail.

As a further study a panel data analysis can be conducted in order to see the evolution of competitiveness of the countries.

References

1. Oral M, Cinar U, Chabchoub H (1999) Linking industrial competitiveness and productivity at the firm level. European Journal of Operational Research 118(2): 271–277
2. WEF (2006) The Global Competitiveness Report 2006–2007. Hampshire: Palgrave Macmillan
3. http://www.imd.ch
4. Sala-i-Martin X, Artadi EV (2004) The Global Competitiveness Index. in [5] pp 51–80

5. WEF (2004) The Global Competitiveness Report 2004–2005. Hampshire: Palgrave Macmillan
6. Onsel Ş, Ulengin F, Ulusoy G, Aktaş E, Kabak Ö, Topcu Yİ (2007) A new perspective on competitiveness of nations. Socio-Economic and Planning Science (in press)
7. Hair J, Anderson KE, Black WC (1995) Multivariate Data Analysis with Readings. Prentice Hall: New York
8. Mangiameli P, Chen SK, West DA (1996) Comparison of SOM neural network and hierarchical clustering. European Journal of Operational Research 93(2): 402–417
9. Kohonen T (1987) Adaptive associative and self-organizing functions in neural computing. Applied Optics 26(23): 4910–4918
10. http://www.mathworks.com
11. http://www.mathworks.com/access/helpdesk/help/dfdoc/nnet/nnet.pdf
12. Yoon Y, Swales G, Margavio TM (1993) A comparison of discriminant analysis versus artificial neural networks. Journal of Operational Research Society 44(1): 51–60
13. Boznar M, Lesjak M, Mlakar P (1993) A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmospheric Environment part B: Urban Atmosphere 27B: 221–230
14. Hwarng HB, Ang HT (2001) A simple neural network for ARMA (p,q) time series. Omega 29: 319–333
15. Swanson NR, White H (1997) Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. International Journal of Forecasting 13(4): 439–461
16. Hruschka H (1993) Determining market response functions by neural network modeling: a comparison to econometric techniques, European Journal of Operational Research 66(1): 27–35
17. Onsel Sahin S, Ulengin F, Ulengin B (2004) A dynamic approach to scenario analysis: the case of Turkey's inflation estimation. European Journal of Operational Research 158(1): 124–145

Chapter 11

Numerical analysis for kinetics and yield of wood biomass pyrolysis

F. Marra

Dipartimento di Ingegneria Chimica e Alimentare, Università degli studi di Salerno, via Ponte Don Melillo 84084 Fisciano, SA, Italy, fmarra@unisa.it

Abstract Solution of ODEs set describing the pyrolysis of wood biomass, on the basis of cellulose, hemicelluloses, and lignin content of investigated materials is proposed as simple but effective tool for estimating the yield of pyrolysis of various biomasses available as by-products of typical agricultural activities in South of Italy. The pyrolysis of wood biomass represents a valid technique for recovering biofuel from residues of forestry and other activities, in agriculture as in industry, where wood and other plant residues are available. Wood biomass is essentially a composite material, the major constituents being cellulose, hemicellulose, lignin, organic extractives, and inorganic minerals. The weight percent of cellulose, hemicellulose, and lignin varies in different species of wood biomass. Results show that hazelnut shells and poplar prunings give high yield in fuel (gas and tar vapors) between 700 and 900 K, both for conventional slow pyrolysis and for fast pyrolysis, whereas olive tree prunings and chestnut wood residues give an appreciable yield (higher than 70%) only for temperatures above 900 K for fast pyrolysis. Sunflower residues, characterized by higher content of non-CHL compounds, give the lower yield in fuel for all the investigated conditions.

Keywords. Biomass, Pyrolysis, CHL model, Rate estimation, Biogas yield, Numerical analysis

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_11,

11.1 Introduction

The development of the pyrolysis process for the biomass conversion and the design of required equipments demand the acquaintance of various aspects: the understanding of the mechanisms governing the process; the acquaintance of the most meant parameters to be estimated during the pyrolysis and their effect on the process; the determination of the devolatilization rate.

In many cases, the description of the pyrolysis rate through relatively simple models is extremely useful from an engineering point of view. In other cases, instead, it is necessary to look at more complex models characterized by an important number of parameters, leading to higher computational costs. It is in the case of models where the pyrolysis of the wood biomass single components (that are essentially, cellulose, hemicellulose, and lignin) is described: in this case one speaks about CHL (Cellu-lose-Hemicellulose-Lignin) model.

The numerous existing studies on the mechanisms of pyrolysis of the biomass and its components [1–5] and on process rate modeling [6–8] have produced various reasonable rate models, all from the analysis of several woody materials and from experiments conducted in a wide range of operating conditions.

The reasons for the different approaches reside in some parameters, such as the particular nature of the processed material; the composition that influences the rates of biomass pyrolysis; the rate of the devolatilization process that depends strongly on the operating conditions, in particular temperature, speed of heating, and time of residence. In fact, there are several important variables, such as the biomass species, chemical and structural composition of the biomass, particle size, temperature, heating rate, atmosphere, pressure, and reactor configuration that affect the yield. In this chapter only the composition of biomass in terms of cellulose, hemicellulose, and lignin (CHL) was considered in order to estimate the pyrolysis yield in volatile compounds, gas and tar vapors, suitable as fuels.

11.2 Problem formulation

The biomass pyrolysis rate was related to wood biomass composition, in order to set up a model useful for all wood materials.

Defining *x*, *y,* and *z* as the percentage of cellulose, hemicellulose, and lignin, respectively, and indicating with v_i, the thermal degradation rate of each of the considered components with respect to its initial mass, one can write

$$v_{BIOMASS} = x \cdot v_C + y \cdot v_H + z \cdot v_L \tag{11.1}$$

and, then, if Y_i is the mass fraction of pyrolysis products with respect to the mass of each component, one can write as follows:

$$Y_{BIOMASS} = x \cdot Y_C + y \cdot Y_H + z \cdot Y_L \tag{11.2}$$

This assumption also considers that all the possible interactions among biomass components have a negligible effect on the advance of the pyrolysis.

In 1979, Bradbury et al. [1] introduced a new model for the description of the rate of wood biomass pyrolysis (it was developed, in particular, for the cellulose); other authors [3, 7] improved the model and extended it to other biomass components (hemicellulose and lignin).

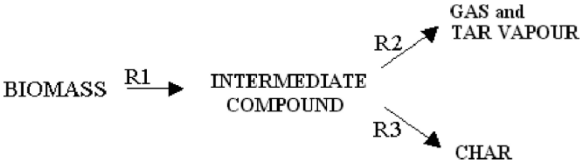


Fig. 11.1. Model of pyrolysis mechanism with intermediate compound

According to Fig. 11.1, the model previews an initial reaction (R1) that describes the global result of all reactions happening below 473 K. Below this temperature, one can observe transformation and a preliminary depolymerization of the starting material, which becomes an active intermediate. This first step is considered as a zero-order rates model [4]. The intermediate active compound is then subjected to further decomposition, following two competitive reactions, leading to gaseous products (gas and tar vapors, as in reaction R2, of order n_2) and char (as in reaction R3, of order n_3).

Equations describing the model are as follows:

$$\frac{dY_B}{dt} = -k_{0,1}\, e^{-\frac{E_{A,1}}{RT}} \tag{11.3}$$

$$\frac{dY_{B^+}}{dt} = k_{0,1}\, e^{-\frac{E_{A,1}}{RT}} - k_{0,2}\, e^{-\frac{E_{A,2}}{RT}} (Y_{B^+})^{n_2} - k_{0,3}\, e^{-\frac{E_{A,3}}{RT}} (Y_{B^+})^{n_3} \tag{11.4}$$

$$\frac{d(Y_G + Y_T)}{dt} = k_{0,2}\, e^{-\frac{E_{A,3}}{RT}} (Y_{B^+})^{n_3} \tag{11.5}$$

$$\frac{dY_C}{dt} = k_{0,3}\, e^{-\frac{E_{A,3}}{RT}} (Y_{B^+})^{n_3} \tag{11.6}$$

where Y_B, Y_{B+}, Y_G, Y_T and Y_C are the mass fraction (that is kg of product with respect to kg of fed biomass) respectively of biomass, intermediate compound, gas, tar vapor and char.

A similar set of equations can be used for each single component: In this way, we do not refer to the whole biomass, but to cellulose, hemicellulose, and lignin.

Initial conditions (*t=0*) are

$$Y_B = 1 \tag{11.7}$$

$$Y_{B^+} = Y_G = Y_T = Y_C = 0 \tag{11.8}$$

In order to evaluate the theoretical yield of pyrolysis products, the biomass is considered to be reduced to particles having a size so small that internal profiles of mass or internal profiles of temperature can be neglected. According to this hypothesis, the linear heating rate of whole biomass is described as follows:

$$T = T_0 + (HR) \cdot t \tag{11.9}$$

where *HR* is the heating rate and T_0 is the initial temperature.

When all the biomass is converted ($Y_B=0$), Eq. (11.3) and the first term in Eq. (11.4) are omitted. In isothermal conditions ($T=T_0$=constant) as well as in non-isothermal ones, Eqs. (11.3), (11.4), (11.5), and (11.6) represent a system of ordinary differential equations, non-linear, that can be solved using a Runge–Kutta algorithm.

11.2.1 Model parameters

Literature [4] states that reactions R2 and R3 are often of order $n_2=n_3$=1.5. In such a case, the fitting of TGA results with results of model for each component allows us to calculate the rate parameters (activation energy, $k_{0,i}$, and frequency factor, $E_{A,i}$), as reported in Table 11.1. These values were used to evaluate the model results in terms is of yield of gas, vapour of tar, and char.

Table 11.1. Rate parameters for CHL model in the temperature range 573 K < T < 873 K [4]

Component	Reaction	n	$k_{0,i}$ [s^{-1}]	$E_{A,i}$ [kJ/mol]
Cellulose	R1	0	2.2e14	167.5
	R2	1.5	9.4e15	216.6
	R3	1.5	3.1e13	196.0
Hemicellulose	R1	0	3.3e6	72.40
	R2	1.5	1.1e14	174.1
	R3	1.5	2.5e13	172.0
Lignin	R1	0	3.3e12	147.7
	R2	1.5	8.6e8	137.1
	R3	1.5	4.4e7	122.1

11.2.2 Materials and methods

As materials for investigation, biomasses available as by-products of typical agricultural activities in South of Italy were considered. Materials and their composition are shown in Table 11.2.

Furthermore, in order to test the hypothesis of an overall reaction order of 1.5, 12 mg samples of beech wood residues (composed of 50.5% cellulose, 29.6% hemicellulose, and 12.7% lignin, being the 7.2% non-CHL components) were subjected to thermogravimetric analysis between 520 and 640 K in a current of nitrogen, using a TA Instruments Q500 TGA. Beech wood residues were used which are available as fine dry powder (average dimension less than 0.08 mm) from a local industry.

Table 11.2. CHL composition for some common biomasses available as by-products of typical agricultural activities in South of Italy

Material	Cellulose	Hemicel.	Lignin	Non-CHL
Hazelnut shells	25.9	29.9	42.5	1.7
Poplar prunings	42.3	31.0	16.2	10.5
Olive tree prunings	22.2	21.1	45.0	11.7
Chestnut wood residues	41.1	16.0	22.8	20.1
Sunflowers residues	27	18	27	28

11.3 Results

Thermogravimetric analysis on 12 mg samples of fine powder obtained from beech wood residues at a heating rate of 20 K s^{-1}, in a current of pure nitrogen gave the results shown in Fig. 11.2. TGA data were used in order to evaluate an overall rate for the pyrolysis reaction.

In fact, for an overall *n*-order rate, the weight decay with the process time is given by the following equation:

$$\frac{dW}{dt} = -k_0 e^{-\frac{E_A}{RT}} W^n \tag{11.10}$$

where *W* is the sample weight and the rate constant is expressed according to the Arrhenius law.

Equation (11.10) can be rearranged in terms of logarithms and it gives the following expression:

$$\ln\left(\frac{-\frac{dW}{dt}}{W^n}\right) = \ln k_0 - \frac{E_A}{RT} \tag{11.11}$$

Guessing the value of *n* using the TGA data, it is possible to plot the first term versus *1/T*: the data will be on a straight line with best correlation factor corresponding to the best value guessed for *n*.

The results of this analysis are reported in Fig. 11.3, where the first term of Eq. (11.11) was plotted against the inverse of temperature. The best correlation is given when *n*=1.5, according to quoted literature [4].

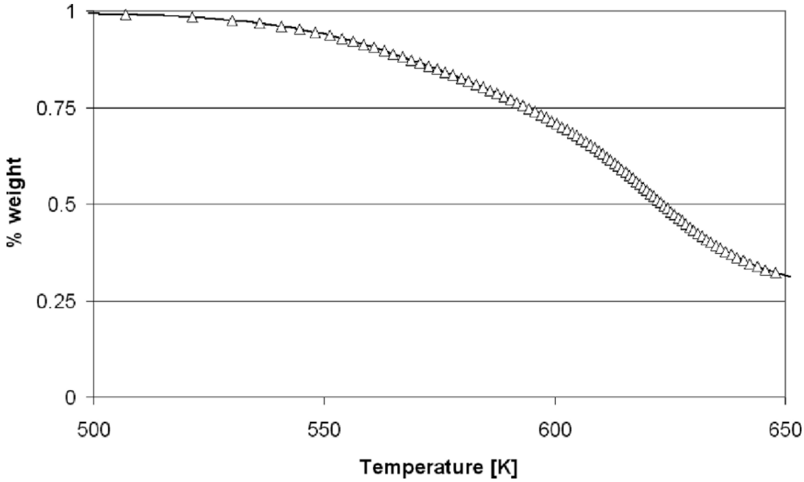


Fig. 11.2. TGA of beech wood residues at a heating rate of 20 K s^{-1}

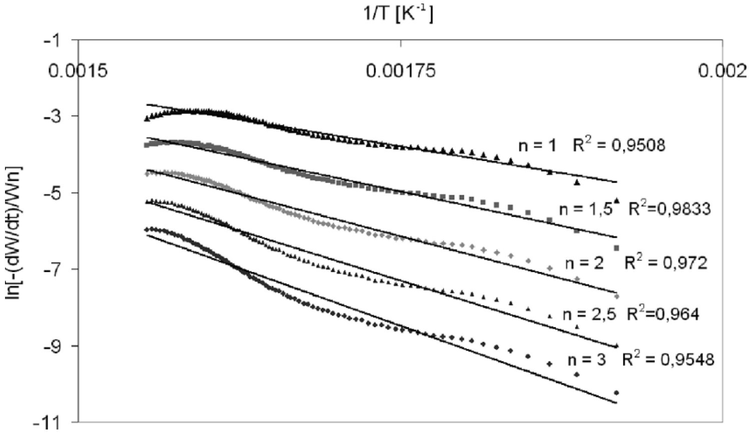


Fig. 11.3. Elaboration of TGA data for the evaluation of overall reaction rate order of beech wood residues show slow pyrolysis

This overall model cannot give any indication about the yield in terms of char and volatile products obtained by the pyrolysis of biomass. For this purpose, the rate model with active intermediate is used.

First of all, theoretical yields of char and gas and tar vapor were calculated for a hypothetical biomass constituted of only cellulose, of only hemicellulose, and of only lignin, respectively.

Results calculated at 700 K are shown in Fig. 11.4. The higher yield of volatile compounds (gas and tar vapour) is given by 100% cellulose biomass, whereas the higher yield of char. is given by 100% lignin biomass.

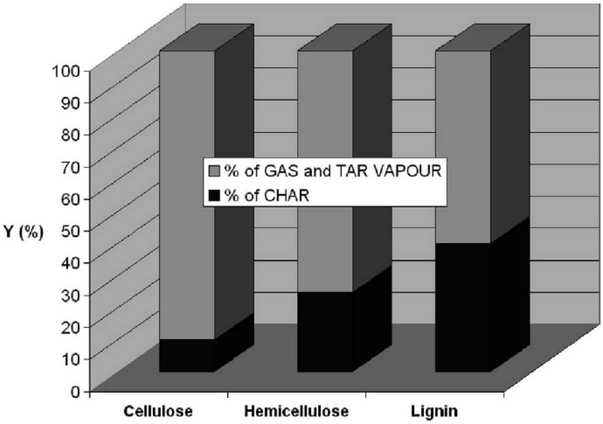


Fig. 11.4. Theoretical yields of char and volatile products of cellulose, hemicellulose, and lignin at 700 K

It is expected that the higher is the content of cellulose with respect to hemicellulose and lignin, the higher the yield in volatile products.

For this purpose, the materials listed in Table 11.2 can be divided into two groups: the first one is composed of poplar prunings and chestnut wood residues, with the higher cellulose content (42.3 and 41.1%, respectively); the second one is composed of hazelnut shells, olive trees prunings, and sunflower residues, with a lower content of cellulose (25.9, 22.2, and 27%).

In the first group, poplar prunings show higher hemicellulose content with respect to the chestnut wood residues, and also a lower content of non-CHL components. In the second group, hazelnut shells and olive tree prunings show similar content of lignin (42.5 and 45%, respectively), but hazelnut shells are characterized by the lowest content of non-CHL components.

Figures 11.5 and 11.6 show the theoretical yields of char and volatile products of considered biomasses, evaluated by solving the CHL model in isothermal conditions at 700 and 900 K, respectively.

Higher yields of volatile products are given by hazelnut shells and poplar prunings for both analyzed temperatures (69% at 700 K and 75% at 900 K).

Higher yields of char are given by hazelnut shells and olive tree prunings, for both analyzed temperatures (25% at 700 K and 18% at 900 K).

Hazelnut shells also show the highest overall yield (almost 95%), whereas the lowest is the one which refers to sunflower residues (almost 71%). Obviously, the overall yield is lower when the content of non-CHL components is higher.

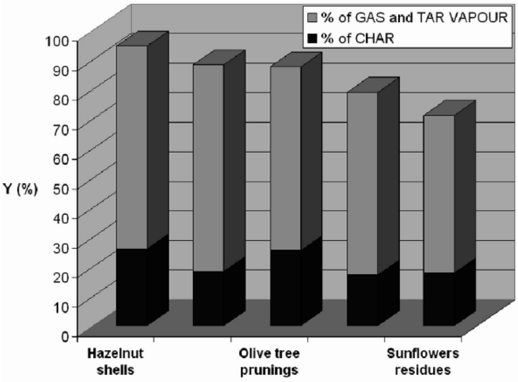


Fig. 11.5. Theoretical yields of char and volatile products of considered biomasses evaluated by solving the CHL model in isothermal conditions at 700 K

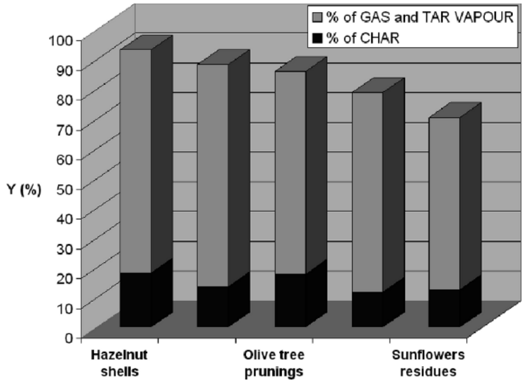


Fig. 11.6. Theoretical yields of char and volatile products of considered biomasses, evaluated by solving the CHL model in isothermal conditions at 900 K

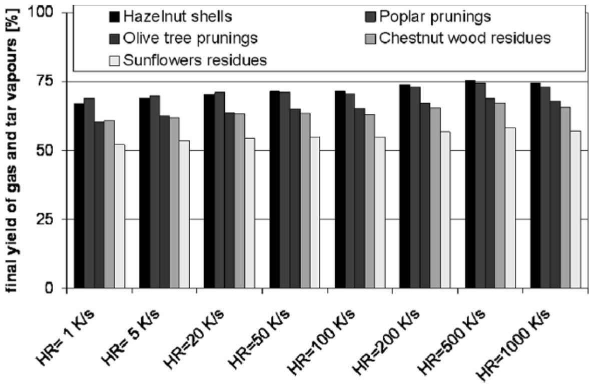


Fig. 11.7. Theoretical yields of gas and tar vapors of considered biomasses, evaluated by solving the CHL model as a function of heating rate (HR)

Figure 11.7 summarizes the theoretical yields of gas and tar vapors obtained solving the model in non-isothermal conditions. During pyrolysis process, in fact, the material is heated in order to reach temperature values between 800 (conventional pyrolysis) and 1000 K (fast pyrolysis). Eight different heating rate (HR) values were considered: four values below 100 K/s, considered as slow or conventional pyrolysis conditions, and four from 100 up to 1000 K/s considered as fast pyrolysis, with residence time between 0.5 and 5 s. Results show that hazelnut shells and poplar prunings give higher yields in gas and tar vapors for all the heating rates investigated; olive tree prunings and chestnut wood residues give an appreciable yield (around 70%) only for temperatures above 900 K for fast pyrolysis conditions. Sunflowers residues, characterized by higher content of non-CHL compounds, give the lower yield in fuel (always less than 60%) for all the investigated conditions. Almost for all considered materials, there is a maximum yield for heating rate of 500 K/s.

11.4 Conclusions

For the estimation of volatile product yield of pyrolysis of wood and plant biomass, a CHL model was presented. Investigation on five biomasses available as by-products of typical agricultural activities in South of Italy allowed us to characterize the role of composition and heating rate on the final yield. Among the considered biomasses, hazelnut shells and poplar prunings gave high yield in fuel (gas and tar vapors) between 700 and 900 K both for conventional slow pyrolysis and fast pyrolysis, whereas olive tree prunings and chestnut wood residues gave an appreciable yield (higher than 70%) only for temperatures above 900 K for fast pyrolysis. Sunflower residues, characterized by higher content of non-CHL compounds, give the lower yield in fuel for all the investigated conditions.

References

1. Bradbury AGW, Sakai Y, Shafizadeh F (1979) A kinetic model for pyrolysis of cellulose. Journal of Applied Polymer Science 23:3271–3280
2. Koufopanos CA, Maschio G, Lucchesi A (1989) Kinetic modeling of the pyrolysis of biomass and biomass components. The Canadian Journal of Chemical Engineering 67:75–84
3. Fisher T, Hajaligol M, Waymack B, Kellogg D (2002) Pyrolysis behavior and kinetics of biomass derived materials. Journal of Analytical and Applied Pyrolysis 62:331–349
4. Di Blasi C (1997) Influences of physical properties on biomass devolatilization characteristics. Fuel 76(10):957–964
5. Thurn F, Mann U (1981) Kinetics investigation of wood pyrolysis. Industrial and Engineering Chemistry Research 20:482–489
6. Antal MJ Jr, Varhegyi G (1995) Cellulose pyrolysis kinetics: The current state of knowledge. Industrial and Engineering Chemistry Research 34(3):703–717
7. Caballero JA, Font R, Marcilla A, Conesa JA (1995) New kinetic model for thermal decomposition of heterogeneous materials. Journal of Applied Polymer Science 34(3):806–812
8. Liang XH, Kozinski JA (2000) Numerical modeling of combustion and pyrolysis of cellulosic biomass in thermogravimetric systems. Fuel 79:1477–1486

Chapter 12

Maintenance of the pre-large trees for record deletion

Chun-Wei Lin[1], Tzung-Pei Hong[2], Wen-Hsiang Lu[1]

[1]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, 701, Taiwan, R.O.C.
{p7895122, whlu}@mail.ncku.edu.tw
[2]Department of Computer Science and Information Engineering, National University of Kaohsiung, Kaohsiung, 811, Taiwan, R.O.C.
tphong@nuk.edu.tw

Abstract. The frequent pattern tree (FP-tree) is an efficient data structure for association-rule mining without generation of candidate itemsets. It, however, needed to process all transactions in a batch way. In addition to record insertion, record deletion is also commonly seen in real applications. In this chapter, we propose the structure of pre-large trees for efficiently handling deletion of records based on the concept of pre-large itemsets. Due to the properties of pre-large concepts, the proposed approach does not need to rescan the original database until a number of records have been deleted. The proposed approach can thus achieve a good execution time for tree construction especially when a small number of records are deleted each time. Experimental results also show that the proposed approach has a good performance for incrementally handling deleted records.

Keywords. Data mining, FP-tree, Pre-large-tree algorithm, Pre-large itemsets, Record deletion

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_12,

12.1 Introduction

Many algorithms for mining association rules from transactions were proposed, most of which were based on the Apriori algorithm [1], which generated and tested candidate itemsets level by level. This may cause iterative database scans and high computational costs. Han et al. proposed the frequent-pattern-tree (FP-tree) structure for efficiently mining association rules without generation of candidate itemsets [2]. Both the Apriori and the FP-tree mining approaches belong to batch mining. That is, they must process all the transactions in a batch way. In real-world applications, new transactions are usually inserted into databases incrementally.

One noticeable incremental mining algorithm was the fast-updated algorithm (called FUP), which was proposed by Cheung et al. [3] for avoiding the shortcomings mentioned above. Although the FUP algorithm could indeed improve mining performance for incrementally growing databases, original databases still needed to be scanned when necessary.

In the past, Hong et al. thus proposed the pre-large concept to further reduce the need for rescanning original database [4]. A pre-large itemset was defined based on two support thresholds. The algorithm did not need to rescan the original database until a number of new transactions had been inserted. Since rescanning the database spent much computation time, the maintenance cost could thus be reduced in the pre-large-itemset algorithm.

Hong et al. also modified the FP-tree structure and designed the fast-updated frequent pattern trees (FUFP trees) to efficiently handle newly inserted transactions based on the FUP concept [5]. The FUFP-tree structure was similar to the FP-tree structure except that the links between parent nodes and their child nodes were bi-directional. Besides, the counts of the sorted frequent items were also kept in the Header_Table of the FP-tree algorithm.

In this chapter, we proposed the structure of pre-large tree for handling the deletion of records based on the concept of pre-large itemsets [4]. A structure of a pre-large tree is to keep not only frequent items but also pre-large items. Based on the pre-large itemsets, the proposed approach can effectively handle cases in which itemsets are small in both an original database and deleted records. The proposed algorithm does not require rescanning the original databases to construct the pre-large tree until a number of deleted records have been processed. Experimental results also show that the proposed algorithm has a good performance for incrementally handling deleted records.

12.2 Review of related works

In this section, some related researches are briefly reviewed. They are the FUFP-tree algorithm and the pre-large-itemset algorithm.

12.2.1 The FUFP-tree algorithm

The FUFP-tree construction algorithm is based on the FP-tree algorithm [2]. The links between parent nodes and their child nodes are, however, bi-directional. Bi-directional linking will help fasten the process of item deletion in the maintenance process. Besides, the counts of the sorted frequent items are also kept in the Header_Table.

An FUFP tree must be built in advance from the original database before new transactions come. When new transactions are added, the FUFP-tree maintenance algorithm will process them to maintain the FUFP tree. It first partitions items into four parts according to whether they are large or small in the original database and in the new transactions. Each part is then processed in its own way. The Header_Table and the FUFP tree are correspondingly updated whenever necessary.

12.2.2 The pre-large-itemsets algorithm

Hong et al. proposed the pre-large concept to reduce the need of rescanning original database [4] for maintaining association rules. A pre-large itemset is not truly large, but may be large with a high probability in the future. Two support thresholds, a lower support threshold and an upper support threshold, are used to realize this concept. The upper support threshold is the same as that used in the conventional mining algorithms. The support ratio of an itemset must be larger than the upper support threshold in order to be considered large. On the other hand, the lower support threshold defines the lowest support ratio for an itemset to be treated as pre-large. An itemset with its support ratio below the lower threshold is thought of as a small itemset. Pre-large itemsets act like buffers and are used to reduce the movements of itemsets directly from large to small and vice versa.

Considering an original database and some records to be deleted by the two support thresholds, itemsets may fall into one of the following nine cases illustrated in Fig. 12.1.

Cases 2, 3, 4, 7, and 8 above will not affect the final association rules. Case 1 may remove some existing association rules, and cases 5, 6, and 9 may add some new association rules. If we retain all large and pre-large itemsets with their counts after each pass, then cases 1, 5, and 6 can be handled easily. Also, in the maintenance phase, the ratio of deleted records to old transactions is usually very small. It has been formally shown that an itemset in case 9 cannot possibly be large for the entire updated database as long as the number of transactions is smaller than the number *f* shown below [4]:

$$f = \left\lfloor \frac{(S_u - S_l)d}{S_u} \right\rfloor,$$

where *f* is the safety number of deleted records, S_u is the upper threshold, S_l is the lower threshold, and *d* is the number of original transactions.

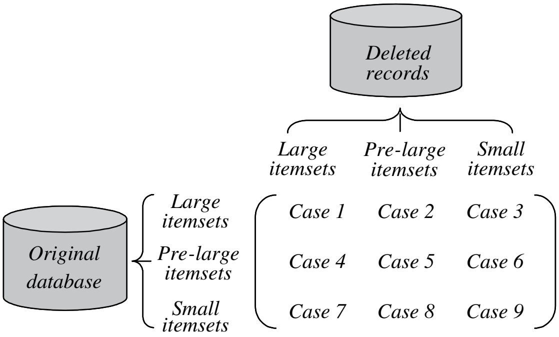


Fig. 12.1. Nine cases arising from and the original database and the deleted records

12.3 The proposed deletion algorithm

A pre-large tree must be built in advance from the initially original database before the records are deleted from the original databases. The database is first scanned to find the large items which have their supports larger than the upper support threshold, and the pre-large items which have their minimum supports lie between the upper and lower support thresholds. Next, the large and the pre-large items are sorted in descending frequencies. The database is then scanned again to construct the pre-large tree according to the sorted order of large and pre-large items. The ordered frequency values of large items and pre-large items are kept in the Header_Table and Pre_Header_Table, respectively. Besides, a variable *c* is used to record the number of deleted records since the last rescan of the original database with d transactions. The details of the proposed algorithm are described below.

The pre-large-tree deletion algorithm:

INPUT: An old database consisting of $(d-c)$ transactions, its corresponding Header_Table and Pre_Header_Table, its corresponding pre-large tree, a lower support threshold S_l, an upper support threshold S_u, and a set of t deleted records.

OUTPUT: A new pre-large tree after records are deleted by using the pre-large tree deletion algorithm.

STEP 1: Calculate the safety number f of deleted records according to the following formula [4]:

$$f = \left\lfloor \frac{(S_u - S_l)d}{S_u} \right\rfloor .$$

STEP 2: Scan the deleted records to get all the items and their counts.

STEP 3: Divide the items in the deleted records into three parts according to whether they are large (appearing in the Header_Table), pre-large (appearing in the Pre_Header_Table), or small (not in the Header_Table or in the Pre_Header_Table) in the original database.

STEP 4: For each item I which is large in the original database, do the following substeps (**Cases 1**, **2,** and **3**):

Substep 4-1: Set the new count $S^U(I)$ of I in the entire updated database as follows:

$$S^U(I) = S^D(I) - S^T(I),$$

where $S^D(I)$ is the count of I in the Header_Table (original database) and $S^T(I)$ is the count of I in the deleted records.

Substep 4-2: If $S^U(I)/(d-c-t) \geq S_u$, update the count of I in the Header_Table as $S^U(I)$ and put I in the set of *Reduced_Items*, which will be further processed in STEP 6;

Otherwise, if $S_u > S^U(I)/(d-c-t) \geq S_l$, remove I from the Header_Table, put I in the head of Pre_Header_Table with its updated frequency $S^D(I)$, and keep I in the set of *Reduced_Item*s;

Otherwise, item I is still small after the database is updated; remove I from the Header_Table and connect each parent node of I directly to its child node in the pre-large tree.

STEP 5: For each item I which is pre-large in the original database, do the following substeps (**Cases 4**, **5,** and **6**):

Substep 5-1: Set the new count $S^U(I)$ of I in the entire updated database as follows:

$$S^{\mathrm{U}}(I) = S^D(I) - S^T(I).$$

Substep 5-2: If $S^U(I)/(d-c-t) \geq S_u$, item I will be large after the database is updated; remove I from the Pre_Header_Table, put I with its new frequency $S^D(I)$ in the end of Header_Table, and put I in the set of *Reduced_Items*;

Otherwise, if $S_u > S^U(\mathrm{I})/(\mathrm{d}-\mathrm{c}-\mathrm{t}) \geq S_l$, item I is still pre-large after the database is updated; update I with its new frequency $S^D(I)$ in the Pre_Header_Table and put I in the set of *Reduced_Items*;

Otherwise, remove item I from the Pre_Header_Table.

STEP 6: For each deleted record with an item J existing in the *Reduced_Items*, subtract 1 from the count of J node at the corresponding branch of the pre-large tree.

STEP 7: For each item I which is neither large nor pre-large in the original database but small in the deleted records (**Cases 9**), put I in the set of *Rescan_Item*s, which is used when rescanning the database in STEP 8 is necessary.

STEP 8: If $t + c \leq f$ or the set of *Rescan_Item*s is *null*, then do nothing;

Otherwise, do the following substeps for each item I in the set of *Rescan_Items*:

Substep 8-1: Rescan the original database to decide the original count $S^D(I)$ of I.

Substep 8-2: Set the new count $S^U(I)$ of I in the entire updated database as follows:

$$S^U(I) = S^D(I) - S^T(I).$$

Substep 8-3: If $S^U(I)/(d-c-t) > S_u$, item I will become large after the database is updated; put I in the set of *Branch_Items* and insert the items in the *Branch_Items* to the end of the Header_Table according to the descending order of their updated counts;

Otherwise, if $S_u > S^U(I)/(d-c-t) \geq S_l$, item I will become pre-large after the database is updated; put I in the set of *Branch_Items*, and insert the items in the *Branch_Items* to the end of the Pre_Header_Table

according to the descending order of their updated counts.
Otherwise, do nothing.

Substep 8-4: For each original transaction with an item J existing in the *Branch_Items*, if J has not been at the corresponding branch of the pre-large tree for the transaction, insert J at the end of the branch and set its count as 1; otherwise, add 1 to the count of the node J.

STEP 9: If $t + c > f$, then set $d = d - t - c$ and set $c = 0$; otherwise, set $c = t + c$.

In STEP 8, a corresponding branch is the branch generated from the large and pre-large items in a transaction and corresponding to the order of items appearing in the Header_Table and the Pre_Header_Table. After STEP 9, the final updated pre-large tree is maintained by the proposed algorithm. The records can then be deleted from the original database.

12.4 An example

In this session, an example is given to illustrate the proposed deletion algorithm for maintaining a pre-large tree when records are deleted. Table 12.1 shows a database to be used in the example. It contains ten transactions and nine items denoted *a–i*.

Table 12.1. The original database in the example

Old database	
TID	Items
1	b, c, e
2	b, c, e, g
3	a, b, d, e, h
4	a, b, e, g, h
5	a, e, g
6	a, b, e
7	b, d, e, g
8	a, b, c, f
9	a, c, d, f
10	c, f, i

Assume the lower support threshold S_l is set at 30% and the upper one S_u at 50%. Here, not only the frequent items are kept in the pre-large tree but also the pre-large items. For the given database, the large items are *b*, *e*, *a,* and *c*, and the pre-large items are *d*, *g,* and *f*, from which the Header_Table and the Pre_Header_Table can be constructed. The pre-large tree is then formed from the database, the Header_Table and the Pre_Header_Table. The results are shown in Fig. 12.2.

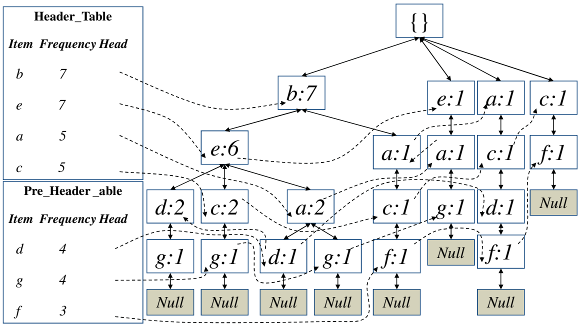


Fig. 12.2. The Header_Table, Pre_Header_Table, and the pre-large tree constructed

STEP 1: The safety number *f* for deleted records is calculated as follows:

$$f = \left\lfloor \frac{(S_u - S_l)d}{S_u} \right\rfloor = \left\lfloor \frac{(0.5 - 0.3)10}{0.5} \right\rfloor = 4.$$

STEP 2: The three records are first scanned to get the items and their counts.

STEP 3: All the items *a* to *i* are divided into three parts, $\{a\}\{b\}\{c\}\{e\}$, $\{d\}\{f\}\{g\}$, and $\{h\}\{i\}$ according to whether they are large (appearing in the Header_Table), pre-large (appearing in the Pre_Header_Table), or small in the original database.

STEP 4: The items in the deleted records which are large in the original database are first processed. In this example, items *a*, *b*, *c*, and *e* (the first partition) satisfy the condition and are processed. Take item *a* as an example to illustrate the substeps. The count of item a in the Header_Table is 5, and its count in the deleted records is 2. The new count of item *a* is thus $5 - 2$ (=3). The new support ratio of item *a* is $3/(10 - 0 - 3)$, which lies between 0.3 and 0.5. Item *a* is removed from the Header_Table and put into the head of the Pre_Header_Table with its updated frequency value and into the set of Reduced_Items. The new count of item *c* is thus $5 - 3$ (=2). The new support ratio of item *c* is $2/(10 - 0 - 3)$, which is lower than 0.3. Item *c* will become small after database is updated. The item *c* is thus removed from the Header_Table and pre-large tree.

The new count of item *b* is 7–1 (=6). Item *b* is thus still a large item after database is updated. The frequency value of item *b* in the Header_Table is thus changed as 6, and item *b* is then put into the set of *Reduced_Items*. Item *e* is similarly processed. After STEP 4, the *Reduced_Items* = {*a*, *b*, *e*}.

STEP 5: The items in the deleted records which are pre-large in the original database are processed. They include items *d*, *f*, and *g*. Take item *d*, *f*, and *g* as an example to illustrate the substeps, respectively. The count of item *d* in the Pre_Header_Table is 4, and its count in the deleted records is 1. The new count of item *d* is thus 4−1 (=3). The new support ratio of item *d* is 3/(10 – 0 – 3), which lies between 0.3 and 0.5. Item *d* is thus still a pre-large item after the database is updated. The frequency value of item *d* in the Pre_Header_Table is thus changed as 4, and item *d* is then put into the set of *Reduced_Items*. The count of item *f* in the Pre_Header_Table is 3, and its count in the deleted records is 3. The new count of item *f* is thus 3 – 3 (= 0). The new support ratio of item *f* is then 0/(10 – 0 – 3), which is smaller than 0.3. Item *f* will become small after database is updated. Item *f* is thus removed from the Pre_Header_Table and from the pre-large tree. The count of item *g* in the Pre_Header_Table is 4, and its count in the deleted records is 0. The new count of item g is thus 4−0 (=4). The new support ratio of item *g* is then 4/(10-0-3), which is larger than 0.5. Item *g* will become large items after database is updated. Item *g* is removed from the Pre_Header_Table and put in the end of Header_Table with its new frequency. The frequency value of item *g* in the Header_Table is thus changed as 4, and item *g* is then put into the set of *Reduced_Items*. After STEP 5, *Reduced_Items* = {*a*, *b*, *d*, *e*, *g*}.

STEP 6: The pre-large tree a updated according to the deleted records with items existing in the *Reduced_Items*. In this example, *Reduced_Items* = {*a*, *b*, *d*, *e*, *g*}. The final results after STEP 6 are shown in Fig. 12.3.

STEP 7: Since the items *h* and *i* are neither large nor pre-large in the original database (not appearing in the Header_Table and in the Pre_Header_Table) and small in the deleted records, it is put into the set of *Rescan_Items*, which is used when rescanning in STEP 7 is required. After STEP 7, *Rescan_Items* = {*h*, *i*}.

STEP 8: Since $t + c = 3 + 0 < f$ (=4), rescanning the original database is unnecessary. Nothing is done in this step.

STEP 9: Since t (= 3) + c (= 0) < f (= 4), set $c = t + c = 3 + 0 = 3$.

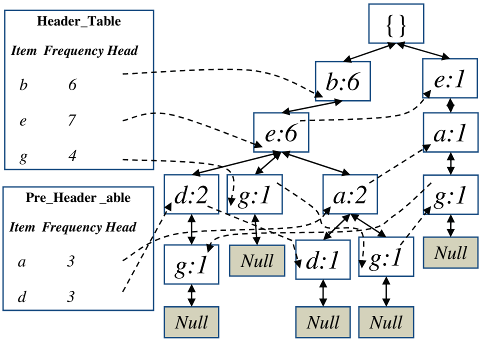


Fig. 12.3. The Header_Table, the Pre_Header_Table, and the pre-large tree after STEP 6

After STEP 9, the pre-large tree is updated. Note that the final value of c is 3 in this example and $f - c = 1$. This means that one more record can be added without rescanning the original database for **Case 9**. Based on the pre-large tree shown in Fig. 12.3, the desired large itemsets can then be found by the FP-growth mining approach as proposed in [2] on only the large items.

12.5 Experiments

Experiments were made to compare the performance of the batch FP-tree construction algorithm, the FUFP-tree deletion algorithm, and the pre-large-tree deletion algorithm for record deletion. The experiments were performed in C++ on an Intel x86 PC with a 3.0 GHz processor and 512 MB main memory and running the Microsoft Windows XP operating system. A real data set called BMS-POS [6] was used in the experiments. The minimum threshold was set at 1 – 5% for the three algorithms, with 1% increment each time. Two thousand transactions were then deleted from the database. For the deletion algorithm of pre-large tree, the upper minimum support threshold was set at 1 – 5% (1% increment each time) and the lower minimum support threshold was set at 0.2, 1.2, 2.2, 3.2, and 4.2%, respectively. The execution times and the number of nodes obtained from the three algorithms were compared. Figure 12.4 shows the execution times of the three algorithms for different threshold values.

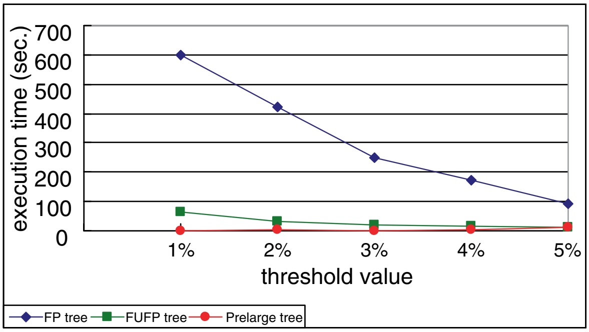


Fig. 12.4. The comparison of the execution times for different threshold values

The comparison of the numbers of nodes for the three algorithms is given in Fig. 12.5.

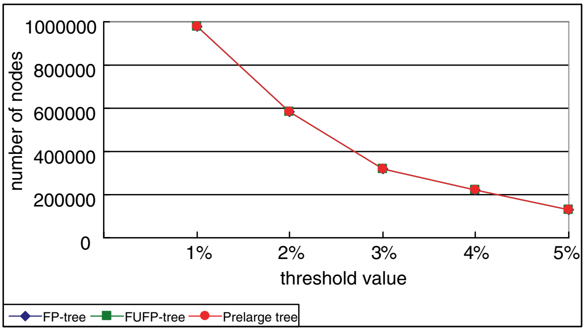


Fig. 12.5. The comparison of the number of nodes for different threshold values

It can be seen that the three algorithms generated nearly the same sizes of trees. The effectiveness of the pre-large tree deletion algorithm is thus acceptable.

12.6 Conclusion

In this chapter, we have proposed the pre-large-tree maintenance algorithm for record deletion based on the concept of pre-large itemsets. Using two user-specified upper and lower support thresholds, the pre-large items act as a gap to avoid small items becoming large in the updated database when transactions are deleted.

Experimental results also show that the proposed pre-large-tree maintenance algorithm runs faster than the batch FP-tree and the FUFP-tree algorithm for handling deleted records and generates nearly the same number of frequent nodes as them. The proposed approach can thus achieve a good trade-off between execution time and tree complexity.

References

1. Agrawal R, Imielinksi T, Swami A (1993) Mining association rules between sets of items in large database. The ACM SIGMOD Conference, pp 207–216
2. Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. The ACM SIGMOD International Conference on Management of Data, pp 1–12
3. Cheung DW, Han J, Ng VT, Wong CY (1996) Maintenance of discovered association rules in large databases: An incremental updating approach. The Twelfth IEEE International Conference on Data Engineering, pp 106–114
4. Hong TP, Wang CY, Tao YH (2001) A new incremental data mining algorithm using pre-large itemsets. Intelligent Data Analysis 5(2):111–129
5. Hong TP, Lin CW, Wu YL (2008) Incrementally fast updated frequent pattern trees. Expert Systems with Applications 34(4):2424–2435
6. Zheng Z, Kohavi R, Mason L (2001) Real world performance of association rule algorithms. The International Conference on Knowledge Discovery and ‘Data Mining, pp 401– 406
7. Mannila H, Toivonen H, Verkamo AI (1994) Efficient algorithm for discovering association rules. The AAAI Workshop on Knowledge Discovery in Databases, pp 181–192

Chapter 13

ALE method in the EFG crystal growth technique

Liliana Braescu[1] Thomas F. George[2]

[1]Department of Computer Science, West University of Timisoara, Blv. V. Parvan 4, Timisoara 300223, Romania
[2]Office of the Chancellor and Center for Nanoscience, Departments of Chemistry & Biochemistry and Physics & Astronomy, University of Missouri-St. Louis, St. Louis, MO 63121, USA

Abstract. Using the arbitrary Lagrangian–Eulerian (ALE) method, an accurate description of the melt flow and impurity distribution in a time-dependent domain $\Omega(t)$, $t\in[0, T]$, is performed. The procedure is developed for a crystal fiber grown from the melt by the edge-defined film-fed growth (EFG) technique, on the basis of the finite-element method using COMSOL multiphysics software. For this, an EFG system without melt replenishment (the melt level in the crucible decreases in time) is considered. By coupling three application modes – incompressible Navier–Stokes, moving mesh arbitrary Lagrangian–Eulerian, and convection–diffusion – it is illustrated, in the time-dependent case, how the pull of the crystal, with a constant rate v_{in}, generates the fluid flow, and it is shown how the resulting fluid flow and deformed geometry determine the impurity distribution in the melt and in the crystal.

Keywords. Arbitrary Lagrangian–Eulerian method, Finite-element technique, Free boundary problems, Edge-defined film-fed growth technique, Single crystal fiber

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_13,

13.1 Introduction

Numerical simulations of viscous incompressible fluid with free boundaries have been receiving more attention over the past few decades. These types of problems arise frequently in several important industrial applications, such as melting and solidification, crystal growth, glass and metal forming processes.

A fundamentally important consideration in developing a computer code for simulating problems of these types is the choice of an appropriate kinematical description of the continuum, which determines the relationship between the deforming continuum and the finite grid or mesh of computing zones and which provides an accurate resolution of material interfaces and mobile boundaries. Algorithms usually use two classical descriptions of motion: the Lagrangian description and the Eulerian description [3, 4]. The Lagrangian description is preferred for "contained fluids" in which there is only small motion or for solid mechanics where the displacements are relatively small, whereas the Eulerian description is preferred for any flow model (except moving boundaries, free surface) in which the mesh would be highly contorted if required to follow the motion. The main disadvantage of the Lagrangian algorithm is its inability to follow large distortions of the computational domain without recourse to frequent remeshing operations. The disadvantages of the Eulerian algorithm are (i) material interfaces lose their sharp definitions as the fluid moves through the mesh and (ii) local regions of fine resolution are difficult to achieve.

A hybrid approach which combines the best features of both the Lagrangian and Eulerian descriptions while minimizing their disadvantages is the arbitrary Lagrangian–Eulerian (ALE) technique [5, 6, 8]. This is the topic of this chapter.

In this chapter, an accurate description of the melt flow and impurity distribution in a time-dependent domain $\Omega(t),\ t\in[0,\ T]$, on the basis of the finite-element method, is performed using COMSOL multiphysics software. Because the geometry (actually the mesh) changes shape, the ALE algorithm is involved. Thus, for determining the impurity distribution in the melt and in the crystal when the fluid domain changes in time, three application modes are coupled in the time-dependent case: incompressible Navier–Stokes (NS), moving mesh ALE, and convection–diffusion (CD). This coupling procedure is developed for a crystal fiber grown from the melt by edge-defined film-fed growth (EFG) technique without melt replenishment, and it demonstrates the ability of COMSOL to simulate flow and concentration evolutions with the help of the moving mesh.

The mathematical model is formulated in two dimensions in a cylindrical-polar coordinate system, attached to the center of the capillary channel, for an Al-doped Si fiber grown from the melt by the EFG method with a central capillary channel shaper (CCC) and without replenishment [1, 2, 7].

13.2 ALE kinematics description

Two configurations are commonly used in continuum mechanics – material configuration and spatial configuration – which are taken as the reference for Lagrangian and Eulerian descriptions. In the ALE description of the motion, neither the material configuration nor the spatial configuration is taken as reference. Thus, a third domain is needed: the referential configuration **P** where reference coordinates m are introduced to identify the grid points (see Fig. 13.1):

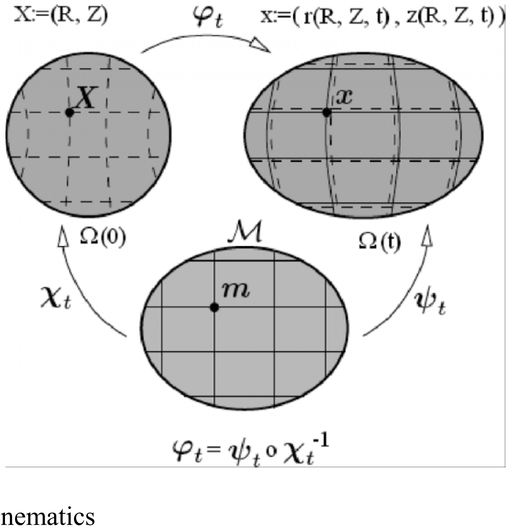


Fig. 13.1. ALE kinematics

We are interested in describing the physical motion between the material configuration $\Omega(0)$ and the spatial configuration $\Omega(t)$, i.e., we want to solve the deformation mapping φ_t: $\Omega(0) \rightarrow \Omega(t)$ for each time $t \in [0, T]$. For this purpose, we consider two additional mappings χ_t: $\Omega(0) \rightarrow$ **P** (material motion) and ψ_t: **P** $\rightarrow \Omega(t)$ (mesh motion). Then the physical motion φ_t may be expressed as $\varphi_t = \psi_t \circ \chi_t^{-1}$, clearly showing that these three mappings are not independent.

Since the choice of the reference configuration is arbitrary, one tries to capture the advantages of both Lagrangian and Eulerian descriptions while minimizing their disadvantages. In particular, the Lagrangian description can be defined as a special case of the ALE description by setting the reference configuration **P** equal to the material configuration Ω(0). Mathematically, this can be obtained by the choice χ_t = *id*, so that the physical motion is equal to the mesh motion, i.e., $\varphi_t = \psi_t$ (see Fig. 13.2).

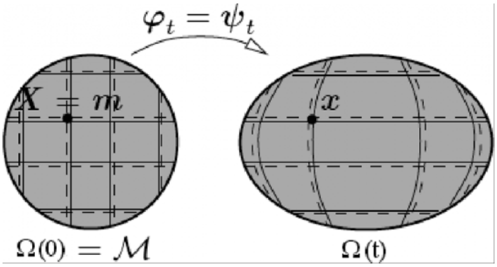


Fig. 13.2. Lagrangian kinematics

Also, the Eulerian description can be defined as a special case of the ALE description. By setting the reference configuration **P** equal to the spatial configuration Ω(*t*) we obtain that the physical motion is equal to the inverse of the material motion $\varphi_t = \chi_t^{-1}$ (see Fig. 13.3).

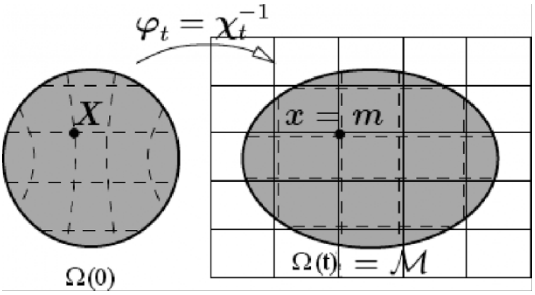


Fig. 13.3. Eulerian kinematics

In the general ALE case, the mapping φ_t: Ω(0)→ Ω(*t*) from the material configuration to the spatial configuration can be understood as the motion of the grid points in the spatial domain. Denoting the velocity of the domain by *w*, in the Eulerian approach *w* is zero, and in the Lagrangian approach *w* is equal to the velocity of the fluid particles. In the ALE approach, *w* is equal to neither zero nor the velocity of the fluid particles, but varies smoothly and arbitrarily between both of them. This arbitrary mesh velocity keeps the movement of the meshes under control according to the physical problem, and it depends on the numerical simulation. More precisely, this method seems to be the Lagrangian description in zones and directions where “small” motion takes place and the Eulerian description in zones and directions where it would not be possible for the mesh to follow the motion of the fluid.

13.3 Mathematical description

The fluid flow and impurity distribution in the crucible, in the capillary channel, and in the meniscus are described in a time-dependent domain $\Omega(t)$, $t\in[0, T]$, by the incompressible Navier–Stokes and conservative convection–diffusion equations,

$$\begin{cases} \rho_l \dfrac{\partial \bar{\mathrm{u}}}{\partial t} + \rho_l (\bar{\mathrm{u}} \cdot \nabla)\bar{\mathrm{u}} = \nabla \cdot \left[-pI + \eta\left(\nabla\bar{\mathrm{u}} + \left(\nabla\bar{\mathrm{u}}\right)^T \right) \right] + \bar{\mathrm{F}} \\ \nabla \cdot \bar{\mathrm{u}} = 0 \\ \dfrac{\partial c}{\partial t} + \nabla \cdot (-D\nabla c + c\bar{\mathrm{u}}) = 0 \end{cases} \tag{13.1}$$

for which axisymmetric solutions are searched in the cylindrical-polar coordinate system (rOz) (see Fig. 13.4). In the system (13.1), $\bar{\mathrm{u}} = (u_r, u_z)$ is the velocity vector, c is the impurity concentration, $\bar{F} = (0, -\rho_l \bar{\mathrm{g}})$ is the volume force field due to gravity, ρ_l is the melt density, p is the pressure, η is the dynamical viscosity, t is the time, and D is the impurity diffusion.

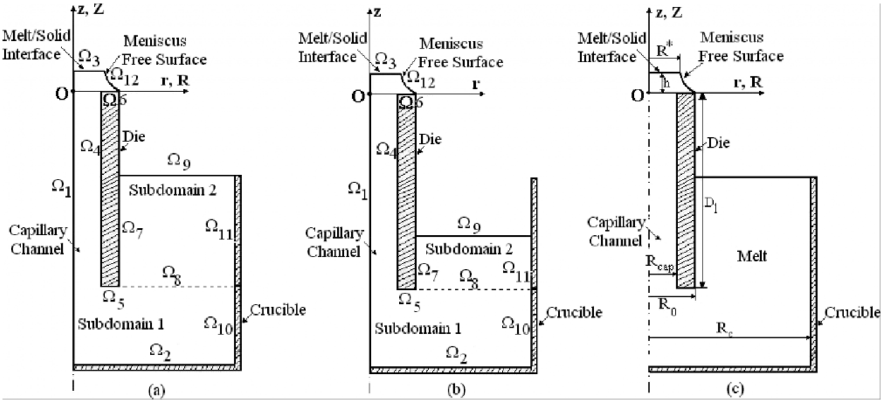


Fig. 13.4. Schematic diagram of the EFG system and boundary regions used in the numerical model

The evolution of

$$\Omega(t) = \{(r(R, Z, t), z(R, Z, t)) \mid (R,Z) \in \Omega(0)\}$$

is described by the system of partial differential equations corresponding to the Laplace smoothing (Poisson equations):

$$\begin{cases} \dfrac{\partial^2}{\partial R^2}\left(\dfrac{\partial r}{\partial t}\right) + \dfrac{\partial^2}{\partial Z^2}\left(\dfrac{\partial r}{\partial t}\right) = 0 \\ \dfrac{\partial^2}{\partial R^2}\left(\dfrac{\partial z}{\partial t}\right) + \dfrac{\partial^2}{\partial Z^2}\left(\dfrac{\partial z}{\partial t}\right) = 0 \end{cases} \tag{13.2}$$

Here *R*, *Z* represent the reference coordinates in the reference frame (*ROZ*), i.e., the fixed frame used for the description of $\Omega(t)$ and of the mesh velocity, as represented in Fig. 13.4(a, b). The solution $(r(R, Z, t), z(R, Z, t))$ satisfies the condition $(r(R, Z, 0), z(R, Z, 0)) = (R, Z)$.

The coupled system (13.1–13.2) is considered in the 2–D domain $\Omega(t)$ with boundaries (Ω_1)–(Ω_{12}), and for solving it, the ALE technique is used. The moving mesh (ALE) application mode solves the system of partial differential equations (13.2) for the mesh displacement. This system smoothly deforms the mesh given by constraints on the boundaries. By the Laplace smoothing option (which has been chosen), the software introduces deformed mesh positions as degrees of freedom in the model.

For solving the coupled system of PDE (13.1–13.2), boundary conditions on Ω_1 to Ω_{12} are imposed, with the *Oz*-axis being considered as a line of symmetry for all field variables:

- Flow (NS) conditions: On the melt/solid interface, the condition of outflow velocity is imposed, i.e., $\overline{\mathrm{u}} = \rho_l / \rho_s \cdot \overline{\mathrm{k}}$, where $\overline{\mathrm{k}}$ represents the unit vector of the *Oz*-axis. On the melt level in the crucible Ω_9 and on the intern boundary Ω_8, we set up the neutral condition, $\left[-\mathrm{pI} + \eta\left(\nabla\overline{\mathrm{u}} + \left(\nabla\overline{\mathrm{u}}\right)^{\mathrm{T}}\right)\right]\overline{\mathrm{n}} = 0$, where $\overline{\mathrm{n}}$ is the normal vector. The other boundaries are set up by the non-slip condition $\overline{\mathrm{u}} = 0$.
- Moving mesh (ALE) conditions: The domain $\Omega(t)$ is divided into two subdomains (labeled 1 and 2 – see Fig. 13.4(a, b)). For subdomain 1, we impose *no displacement* (i.e., subdomain 1 is fixed), and for subdomain 2 we impose *free displacement* (is free to move). Hence, the mesh displacement takes place only in subdomain 2, and it is constrained by the boundary conditions on the surrounding boundaries Ω_7, Ω_8, Ω_9, and Ω_{11}. The displacement in subdomain 2 is obtained by solving the system (13.2). The boundary conditions involve variables from the NS

application mode. To obtain convergence, it is important for the boundary conditions to be consistent. The usual point-wise constraints or ideal constraints for ALE cause unwanted modifications of the boundary condition for the other two application modes (NS and CD). For this reason, in the ALE application mode, we must use non-ideal weak constraints on the boundaries:

- the mesh displacements in the *r*-direction and *z*-direction on Ω_9 are *dr* = 0*,* *dz* = $v_n \times t$*,* where $v_n = -R^{*2} / R_c^2 - R_0^2 \cdot v_{in}$ (the mesh velocity should be equal to the fluid velocity, according to [5, 6, 8]);
- the mesh displacement in the *r*-direction on $\Omega_7 U \Omega_{11}$ is *dr* = 0; the mesh displacement *dz* in the *z*-direction is not specified, i.e., the mesh will follow the fluid movement;
- the mesh displacements *dr* and *dz* in the *r*-direction and *z*-direction are not specified on Ω_8 (the mesh follows the fluid flow).

- Concentration (CD) conditions: On the melt/solid interface, the flux condition is imposed, which expresses that impurities are rejected back into the melt according to $\frac{\partial c}{\partial \overline{\mathrm{n}}} = -\frac{v_{in}}{D}(1 - K_0)c$. On the inner boundary Ω_8, we set up the continuity condition (flux difference is zero), $\overline{\mathrm{n}} \cdot (\mathrm{N}_1 - \mathrm{N}_2) = 0,\ \mathrm{N}_i = -D_i \nabla c_i + c_i \overline{\mathrm{u}}_\mathrm{i}$, where *i* = 1 for subdomain 1 and *i* = 2 for subdomain 2. The other boundaries are set up by the no-impurity flux condition, i.e., insulation: $\frac{\partial \mathrm{c}}{\partial \overline{\mathrm{n}}} = \overline{\mathrm{n}} \cdot \nabla \mathrm{c} = 0 \cdot$

Besides the above boundary conditions, we have to add the initial conditions:

- for the fluid flow: $u(t_0) = 0$*,* $v(t_0) = 0$ in subdomain 1 and $u(t_0) = 0$*,* $v(t_0) = v_\mathrm{n}$ in subdomain 2 (the fluid velocity should be equal to the mesh velocity, according to [5, 6, 8]);
- for the pressure: $p(t_0) = P_0 = -\rho_l \cdot g \cdot z$ in subdomain 1 and $p(t_0) = 0$ in subdomain 2;
- for the initial impurity distribution: $c = C_0 = c(t_0)$ in both subdomains;
- for the mesh displacement: $r(t_0)$=R*,* $z(t_0)$=Z.

Details concerning the significance of these quantities and their values for the Al-doped Si rod are presented in Table 13.1 and Fig. 13.4(c).

Table 13.1. Material parameters for silicon

Nomenclature		Value
c	impurity concentration (mol/m^3)	
C_0	alloy concentration (at %)	0.01
D	impurity diffusion (m^2/s)	5.3×10^{-8}
D_l	die length (m)	0.04
g	gravitational acceleration (m/s^2)	9.81
h	meniscus height (m)	0.5×10^{-3}
K_0	partition coefficient	0.002
η	dynamical viscosity (kg/m×s)	7×10^{-4}
p	pressure (Pa)	0
R	crystal radius (m)	1.5×10^{-3}
R_{cap}	capillary channel radius (m)	1.5×10^{-3}
R_c	inner radius of the crucible (m)	23×10^{-3}
R_0	die radius (m)	2×10^{-3}
ρ_l	density of the melt (kg/m^3)	2500
ρ_s	density of the crystal (kg/m^3)	2300
$\overline{\text{u}}$	velocity vector	
v_{in}	pulling rate (m/s)	10^{-6}
z	coordinate in the pulling direction	

13.4 Numerical results

Numerical investigations are carried out for an Al-doped Si rod of radius $R^* = 1.5\times10^{-3}$ m, grown in terrestrial conditions with a pulling rate $v_{in} = 10^{-6}$ m/s. The boundaries presented in Fig. 13.4 are determined from the particularities (characteristic elements) of the considered EFG growth system. Thus, a crucible with inner radius $R_c = 23\times10^{-3}$ m is considered in which a die of radius $R_0 = 2\times10^{-3}$ m and length 40×10^{-3} m is introduced, such that 2/3 of the die is immersed in the melt. In the die, a capillary channel is manufactured with a radius $R_{cap} = 1.5\times10^{-3}$ m, through which the melt climbs up to the top of the die. A seed is placed in contact with the melt; due to the heat transfer, this seed melts, and a low meniscus with height $h = 0.5\times10^{-3}$ m is developed. These above values define the initial geometry $\Omega(0)$ in the fixed reference frame (ROZ), at $t = 0$.

We then start the growth process of the Al-doped Si rod with a constant pulling rate v_{in}. Because the EFG system is without melt replenishment (the melt level in the crucible decreases in time) and the crystal is pulled with a constant rate v_{in} (at the boundary Ω_3), a fluid flow in the crucible is induced. Hence, the melt height in the crucible decreases, i.e., the length of the boundaries Ω_7, Ω_{11} decreases and the boundary Ω_9 goes down. In this way, the initial geometry $\Omega(0)$ changes in time as $\Omega(t) =\{(r(R, Z, t), z(R, Z, t)) \mid (R,Z)\in \Omega(0)\}$ with respect to the reference frame (*ROZ*). The new reference frame will be the one determined by the spatial coordinates *r, z* of the ALE-frame in which the mesh is moving, i.e., $(r, z)\in \Omega(t)$.

In order to evaluate in which way the resulting fluid flow and the deformed geometry determine the impurity distribution in the melt and in the crystal, using COMSOL multiphysics 3.2 software, we solve the coupled NS-ALE-CD application modes in the ALE-frame (*rOz*). COMSOL multiphysics solves the math necessary to manipulate, move, and deform the mesh simultaneously with the boundary movement, as required by the other coupled physics (see Fig. 13.5).

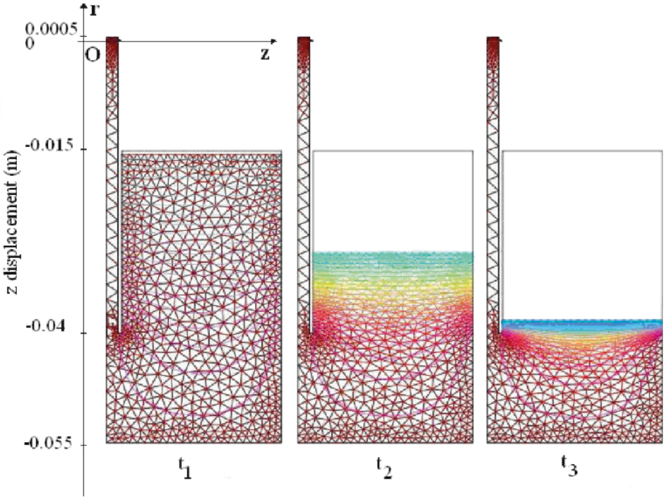


Fig. 13.5. Mesh deformation in the ALE application mode for $v_n = -4.65\times10^{-8}$ m/s at three different instants of time, $0 < t_1 < t_2 < t_3 <T$

The employed mesh is considered with maximum element size of 1e-3 and manually refined along the boundaries 2, 6, 7, 8, 9, 10, 11, and 12 (maximum element size is 1e-4), i.e., along the free surfaces and their neighborhood boundaries. According to the considered geometry, 1703 triangular mesh elements are used. By the NS-CD-ALE equations and Laplace smoothing option software, 23,397 degrees of freedom are introduced for these meshes.

The dependence of the impurity distribution on the geometry changes is presented in Fig. 13.6(a– c).

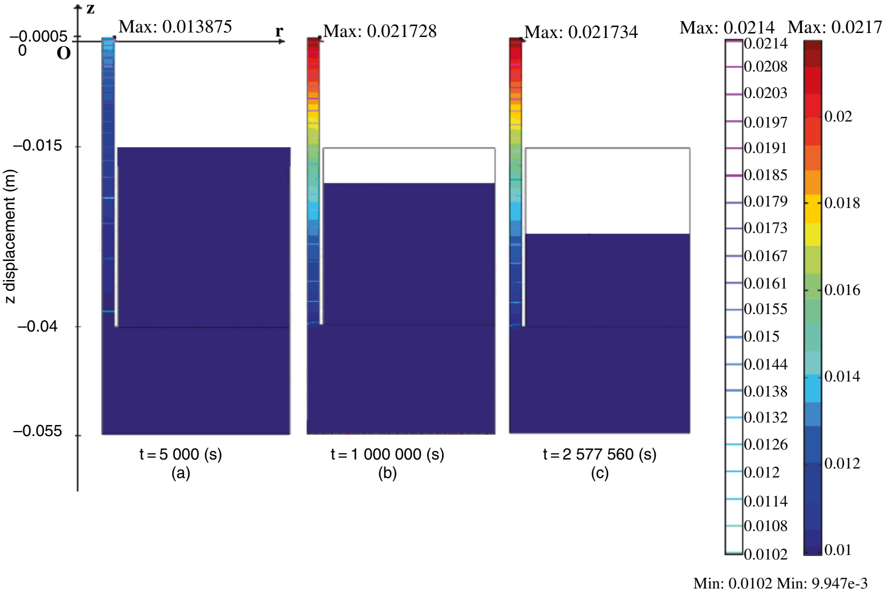


Fig. 13.6. Dependence of the impurity distribution on the decrease of the melt level in the crucible

Computations show that at the beginning, if the melt level decreases in the crucible, then the concentration increases starting from the initial value $C_0 = 0.01$ mol/m^3. Moreover, there exists a certain time after which the impurity concentration becomes constant, even if the melt level still decreases in the crucible.

13.5 Conclusions

The arbitrary Lagrangian–Eulerian (ALE) method for coupled Navier–Stokes and convection–diffusion equations with moving boundaries has been implemented for a melting–solidification process. The effect of the deformed geometry on the impurity distribution has been studied for an Al-doped Si fiber grown from the melt by the edge-defined film-fed growth (EFG) technique.

Acknowledgment. We are grateful to the North Atlantic Treaty Organisation (Grant CBP.EAP.CLG 982530) for support of this project.

References

1. Bunoiu O, Nicoara I, Santailler JL, Duffar T (2005) Fluid flow and solute segregation in EFG crystal growth process. Journal of Crystal Growth 275:e799–e805
2. Braescu L, Balint S, Tanasie L (2006) Numerical studies concerning the dependence of the impurity distribution on the pulling rate and on the radius of the capillary channel in the case of a thin rod grown from the melt by edge-defined film-fed growth (EFG) method. Journal of Crystal Growth 291:52–59
3. Donea J, Fasoli-Stella P, Giuliani S (1976) Finite element solution of the transient fluid-structure problem in Lagrangian coordinates. Proceedings of the International Meeting on Fast Reactor Safety and Related Physics, Chicago 3:1427–1435
4. Donea J, Fasoli-Stella P, Giuliani S (1977) Lagrangian and Eulerian finite-element techniques for transient fluid-structure interaction problems. Trans. SMiRT-4, San Francisco
5. Duarte F, Gormaz R, Natesan S (2004) Arbitrary Lagrangian-Eulerian method for Navier-Stokes equations with moving boundaries. Computer Methods in Applied Mechanics and Engineering 193:4819–4836
6. Fernandez M, Moubachir M (2002) Sensitivity analysis for an incompressible aeroelastic system. Mathematical Models and Methods in Applied Sciences 12:1109–1130
7. George TF, Balint S, Braescu L (2007) Mass and heat transport in Bridgman-Stockbarger and edge-defined film-fed growth systems. Springer Handbook of Crystal Growth, Defects and Characterization, Springer, Berlin Heidelberg New York
8. Hughes TJR, Liu WK, Zimmermann TK (1981) Lagrangian-Eulerian finite-element formulation for incompressible viscous flows. Computer Methods in Applied Mechanics and Engineering 29:329–349

Chapter 14

Application of e-learning methods in the curricula of the faculty of computer science

J. P. Nowacki[1], L. Banachowski[2]

[1] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, POLAND, jnowacki@pjwstk.edu.pl
[2] Polish-Japanese Institute of Information Technology, Koszykowa 86, 02-008 Warsaw, POLAND, lech@pjwstk.edu.pl

Abstract. The purpose of this chapter is to present curricula development at Polish-Japanese Institute of Information Technology (PJIIT) embracing e-learning methods. We will discuss the current state of development and the conclusions which can be drawn from the 5 years of experience of running online courses at PJIIT. We develop further methods described earlier in [1–3].

In the first part we will describe the methodology we use and show PJIIT's e-learning platform EDU, developed by our own specialists. We will also discuss the specific issues connected with undergraduate, graduate, and postgraduate curricula offered by PJIIT. One of the key features of education at PJIIT is the possibility of choosing between stationary and distance way of learning a subject.

We will also present the use of the EDU system in teaching stationary courses because both the methodology and the technology can be applied not only in distance education but also to enhance traditional studies as well. It seems that the differences in stationary and distance education are not becoming more pronounced with the development of new technologies.

In the second part of this chapter we present basic assumptions of the UNDP project aiming at establishing similar curricula in Ukraine in cooperation with local universities. The project, entitled "Transfer of Information Technology to Ukraine" was coordinated by Polish-Japanese Institute of Information Technology under the

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_14,

auspices of United Nations Development Program (UNDP) and financed with the funds donated by the government of Japan. It began in mid-2004 and was completed in December 2006. This extends the ideas presented earlier in [4] and [5].

Keywords. Online learning, Distance learning, Blended learning, Learning management system, Platform EDU, International e-learning projects, Curricula, Methodology

14.1 Introduction

Polish-Japanese Institute of Information Technology (PJIIT) is the leading Polish university specializing in computer science. It was founded in 1994 as a result of the agreement between the governments of Poland and Japan. The institute offers undergraduate, graduate, and postgraduate courses in the main fields of computer science.

Over the last 8 years the lectures at PJIIT have been gradually moved to electronic format (mostly PowerPoint). What is more, the Internet has become a powerful medium of communication between students and faculty. We have arrived at a conclusion that the time has come to introduce the new form of studies based on the Internet in addition to regular stationary courses. In the year 2001 we started teaching online courses on an experimental basis in cooperation with the University of North Carolina, Charlotte. The participants of the courses were graduate students from both the universities.

In June 2000 the Senate of PJIIT took a decision to start preparation for online studies toward B.Sc. degree in computer science. The new studies commenced in September 2002 with 42 students enrolled. In 2006 they were extended with the studies toward M.Sc. degree in computer science. Their curriculum comprises basic knowledge in computer science enabling our graduates to find jobs as administrators, analysts, system designers, programmers, multimedia, or network designers. Our graduates should possess the ability to set up information systems running over the Internet and the ability to work and cooperate over the Internet. Building the last two skills is inherent in the nature of the courses themselves, i.e., being conducted online. We believe this is the advantage e-studies have over other forms of tuition and we are convinced this advantage needs to be explored. Moreover, after completing studies toward M.Sc. degree in computer science our graduates can design e-learning solutions for higher education and business and do research in the area of IT and its applications.

The curriculum of the new studies is almost identical to the curriculum of stationary studies toward the B.Sc. and M.Sc. degrees in computer science at the Polish-Japanese Institute of Information Technology. The only difference is greater emphasis placed on Internet technologies. In particular, the course "Application of e-learning in higher schools and business" is required for online students and optional for others.

From October 2008 we plan to start our online postgraduate studies in computer science. The studies will be offered for IT specialists as well as specialists in other domains who want to apply IT solutions supporting their everyday activities at work, specifically in conducting software projects. The ideal candidates are the persons who possess basic IT knowledge and have some experience in using IT tools and who want to enhance their knowledge and abilities in the area of design and build information systems, applications, and databases. These studies are supposed to implement the idea of lifelong learning.

Each online student has to come to the institute for 1-week stationary sessions two or three times a year. During these visits they take examinations and participate in laboratory courses requiring specialized equipment. The new studies are based on the educational, multimedia materials available both online and on CDs produced and supplied by PJIIT.

The courses run either exclusively over the Internet or in the mixed mode: lectures over the Internet and laboratory classes at the institute's premises. Each course comprises 15 units treated as lectures. The content of one lecture is mastered by students during 1 week. At the end of the week the students send the assignments to the instructor and carry out tests, which are automatically checked and graded by the system. The grades are entered into the gradebook – each student can see only his or her own grades.

Besides home assignments and tests there are online office hours held 2 hours a week; seminars and live class discussions. Bulletin boards, timetables, discussion forums, and FAQ lists are also available.

It is important that during their studies the students have remote access to the PJIIT's resources such as software, applications, databases, an ftp server, an e-mail server.

Partial grades obtained during the semester (coming mainly from home assignments and tests) contribute to the final grade for online classes. Of course, besides this grade we have always the second grade resulting from the examination administered in the PJIIT building.

It should be emphasized that in our form of studies the attendance of lectures and classes has been replaced by the necessity of systematic, week-by-week, individual work. An online student has to be more responsible, systematic, and self-driven than a stationary student. Every week they are required to demonstrate the understanding of a part of the material by doing homework assignments and tests. It is an exacting form of studies, difficult for many students as the analysis of students' performance shows. Those students who survive the first semester, they number about 50%, seem to learn endurance and systematic work and they are able to continue their hard work during the next semesters.

14.2 Format of lecture materials

Course materials used in the online studies take the form of electronic textbooks in the HTML format, with navigation implemented by means of scripts of JavaScript language. They include (1) material which during traditional lectures is normally displayed on the screen, usually consisting of text and figures; (2) explanations given by the lecturers during their lectures; (3) multimedia presentations whose aim is to help the students to understand the most difficult parts of the course; (4) auxiliary material such as glossary of terms and index; (5) bibliography and webgraphy; (6) references to other materials; (7) homework assignments to be done by students and short questions embedded in the text. For each course there is a syllabus presenting its most important points such as a description of course content, prerequisites, obligatory and recommended textbooks, grading policy.

The form of course materials described above is easily transferable to arbitrary learning environments whose lecture materials are of the form of HTML documents with navigation based on JavaScript language.

14.3 Help in creating virtual community

We have found out that the students who contact one another and the lecturers perform much better. Therefore, before starting online studies, we try to help in creating virtual community between students and faculty. After coming to the institute for the first time, the new students attend the meeting with teachers at the institute's premises. They spend their first week at the institute's premises on a 5-day course devoted to working and collaborating on the Internet. Their first totally online course is optional and concerns High School Math with two aims in mind: equalizing the level of mathematical preparation and providing exercises in online learning before the basic courses start.

14.4 Internet-based system for online studies management

System EDU was created by PJIIT students and staff. The software was written in the programming languages Visual Basic, Java, and ASP. The data are stored in the Microsoft SQL Server database.

The system consists of the following modules:

Main Page – The module enabling presentation of the list of courses to which the user has been granted access and all the new information concerning these courses (such as new announcements).

Course – The module enabling the instructor to present basic information about the course such as the instructor's e-mail address, brief course description, and selection of modules to be used by students during the course. The instructor decides whether the course is disabled or enabled to the students (for example, disabling the course when major modifications take place). The module includes the display of the list of the students of the course with their e-mail and web page addresses. The course participants including the instructor can send e-mail messages to groups of students by selecting their e-mail addresses.

Calendar – Before starting the course, the instructor prepares the course calendar in the form of a list of events each accompanied by a date. The events are the deadlines for homework assignments, tests, dates of online seminars, and so on.

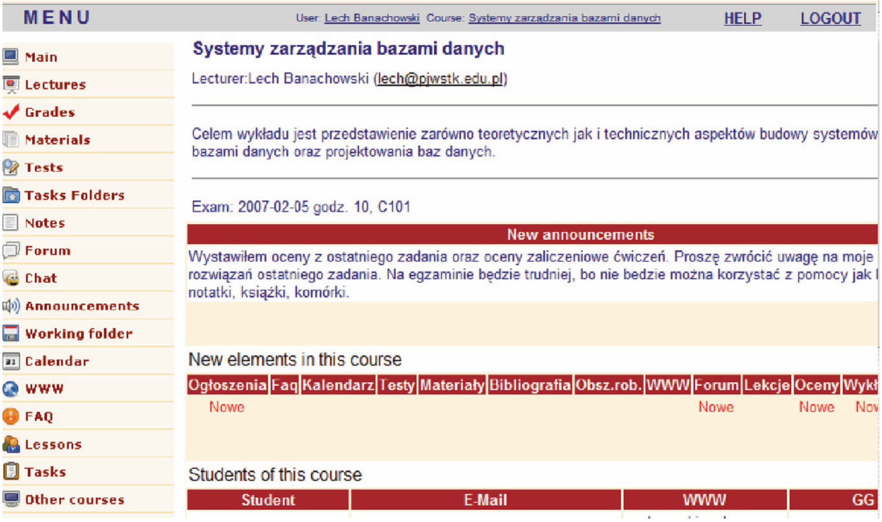
Announcements – The instructor can publish current announcements concerning the course. Announcements are sorted by date. New announcements (last 30 days) are also shown on the Main Page in the section "New announcements."

WWW – The module is a collection of links to WWW pages concerning the course – including pages presenting the instructor and all the students in the course.

Chat – The module enables direct online communication among students and instructors. The instructor can carry out office hours, seminars, and lectures online. On specific dates students and instructor enter the class chatroom and conduct online sessions. On the right-hand side of the window they can see the list of all the participants who have entered the chatroom. Every participant can write text messages visible to all in the chatroom.

In addition, students and instructors can use whiteboards

- to draw images,
- to display e-content stored in files including text and graphics.



Systemy zarządzania bazami danych

Lecturer:Lech Banachowski (lech@pjwstk.edu.pl)

Celem wykładu jest przedstawienie zarówno teoretycznych jak i technicznych aspektów budowy systemów bazami danych oraz projektowania baz danych.

Exam: 2007-02-05 godz. 10, C101

New announcements

Wystawiłem oceny z ostatniego zadania oraz oceny zaliczeniowe ćwiczeń. Proszę zwrócić uwagę na moje rozwiązań ostatniego zadania. Na egzaminie będzie trudniej, bo nie bedzie można korzystać z pomocy jak l notatki, książki, komórki.

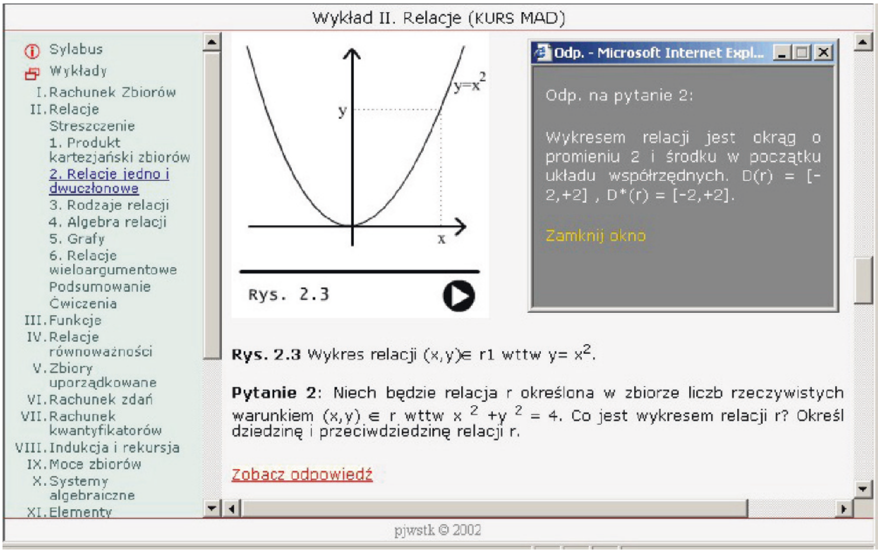
New elements in this course

Ogłoszenia	Faq	Kalendarz	Testy	Materiały	Bibliografia	Obsz.rob.	WWW	Forum	Lekcje	Oceny	Wykł
Nowe								Nowe		Nowe	Nov

Students of this course

Student	E-Mail	WWW	GG

Fig. 14.1. Main course page of EDU



Odp. na pytanie 2:

Wykresem relacji jest okrąg o promieniu 2 i środku w początku układu współrzędnych. D(r) = [-2,+2] , D*(r) = [-2,+2].

Zamknij okno

Rys. 2.3 Wykres relacji (x,y)∈ r1 wttw y= x^2.

Pytanie 2: Niech będzie relacja r określona w zbiorze liczb rzeczywistych warunkiem (x,y) ∈ r wttw x ^{2} +y ^{2} = 4. Co jest wykresem relacji r? Określ dziedzinę i przeciwdziedzinę relacji r.

Zobacz odpowiedź

Fig. 14.2. Chat module

The instructor can restrict the number of students who can enter a specific chatroom, making possible private consultations.

Forum – The module enables off-line communication among students and instructors. Forum is organized by main topics or threads, called "threaded discussions." The instructor can carry out lessons in the form of discussions on selected topics.

FAQ – Frequently asked questions is the list of questions and problems frequently asked by students (e.g., by e-mail or forum) with explanations given by the instructor.

Tasks Folders, Working Folder – The areas for exchanging files among students and the lecturer. There are two types of folders: *Working Folder* – files in this folder may be downloaded by every user in the course and *Tasks Folders* – files in this folder may be uploaded by students, but only the instructor has a privilege to download them – the folder is mainly used to collect homework assignments.

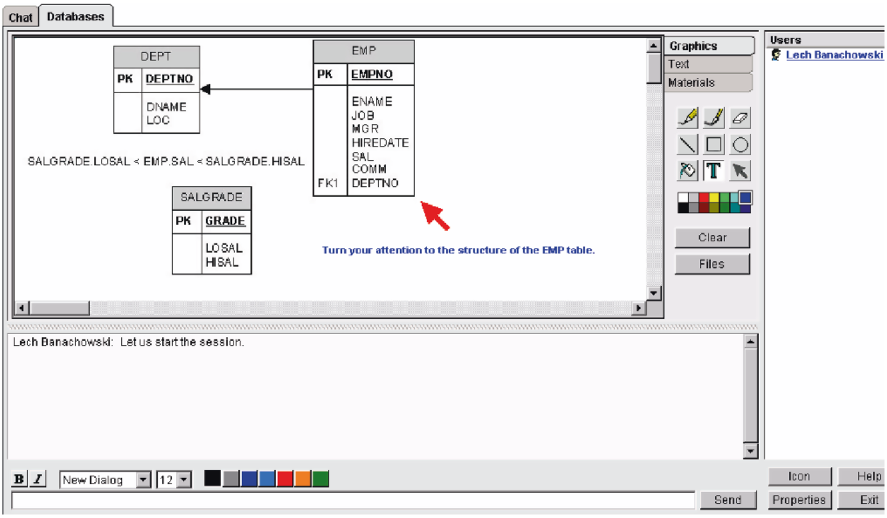
Lectures – Each lecture is an HTML presentation including graphics and multimedia all combined into one learning structure. System EDU makes it easy for a lecturer to make changes in the files forming the lecture presentation. When the student enters the module, a new browser window is opened with the lecture presentation inside.

Materials – Educational materials provided by the instructor for the students to help them in their studies. System EDU does not restrict type of materials. It may by any file. If a student selects a file, which is saved in format recognizable by his or her browser (e.g., html), then the material will be opened in a browser window. But if it is of unrecognizable format, then students will be asked if they want to save the file or open it from current location.

Textbooks – This is the list of textbooks recommended by the instructor of the course. The information goes with title, author, publisher, and description. Attribute "Main" determines if the book is obligatory or supplementary for the course.

Tests – Using this module the instructor prepares tests to be taken by students. Every test may be taken only once (it counts when the student saves the test so he or she can first get acquainted with the test questions and, possibly, during another session provide solutions to them). There are five types of test questions: text, multi-line text, yes/no, options (only one answer is true), and multi-options (there may be more than one answer true).

Grades – The instructor enters the students' grades received for doing specific tasks such as homework assignments, tests, discussions, projects. The instructor can attach additional comments and remarks explaining observed achievements, errors, and shortcomings. The student can see only his or her grades and additionally, how he or she stands in the ranking in comparison to other students. The instructor can print course protocols automatically filled with final grades.



Turn your attention to the structure of the EMP table.

Lech Banachowski: Let us start the session.

Fig. 14.3. Lecture presentation

Lessons – In this module the instructor can design a programed course consisting of interactive lessons. The student studies the programed course by reading materials and answering test questions after the end of each section, each lesson, and the whole course. The student moves to the next lesson only after getting enough points for the test ending the previous lesson. The instructor decides about the number of points to pass the lesson and the whole programed course, respectively. The default threshold is set at 50%.

Notes – In this module users can enter and store text notes. Every note is visible only to its author in the current course along with the date of its creation.

Register – Every entry to a course in the EDU system is stored in the database. Using this module, the instructor can monitor how many times each student used the system and can find the date and time of every visit. Moreover the instructor can display the number of visits at each module for each student.

14.5 Analysis of using the EDU system

All entries to the courses and their modules in the EDU system are registered. Therefore it is possible to monitor the activity and performance of students. In particular, by analyzing the number of entries, we can get information about students who ceased to participate actively in the online course. The following conclusions can be drawn:

1. The average student enters each course 3–5 times a week.
2. The students use the test module most frequently, often look into their grades, and often read lectures. Because the students get their lectures, on the CDs as well, the statistics does not reflect fully the students sinterest in the content.
3. The students who use the system regularly perform much better than others.

14.6 Significance of the new form of studies for the Institute as a whole

The introduction of Internet-based studies has had a stimulating impact on the whole institute. As a side effect the textbooks based on lecture materials originally prepared for the online studies are published by our institute's publishing unit. It is regarded as the most valuable publishing initiative in computer science in Poland. We have observed that the electronic materials prepared for the online studies exert influence on the way ordinary lectures are presented – improving greatly their quality and visual attractiveness. The automatic system of tests prepared for the online studies is used also to carry out ordinary examinations for stationary courses. Moreover, the stationary students who have not passed a course can repeat it enrolling in the online mode, without the need of repetition of classes in the PJIIT building. It has become a popular form of catching up. Another interesting phenomenon is that persons who stopped their studies at the PJIIT several years ago come back and resume their studies as Internet students.

14.7 Using online methods for stationary courses

The differences between stationary and non-stationary studies seem to diminish along with the development of online learning methods. We have found the following valuable elements of electronic methods for stationary courses: possibility of online studying of lectures instead of coming to attend lectures at the university, enabling computer-aided individual work of students, diversifying contacts with teachers and other students.

The following Edu modules are used for stationary courses: Announcements, Task Folders, Working Folders, Lecture materials, Tests/exams at the Institute premises, Discussion Forum, Chat.

14.8 Conclusions

Summing up, in our version of internet-based studies we teach most of the courses exclusively over the Internet. Such courses include among others programming, software engineering, networks, databases, and mathematics. only few subjects require highly specialized technical equipment and such courses are organized in the PJIIT laboratories. These courses include computer graphics, multimedia, and electronics. Internet-based studies have been positively evaluated by most of our students – such a conclusion can be drawn from our standard evaluation check carried every semester. We have noticed that such a form of studies is of particular importance for those students who either stay abroad a lot or who cannot participate regularly in stationary classes because of health, job, or family reasons.

14.9 UNDP project

The UNDP project was sponsored by the Japanese Government, managed by UNDP, and conducted by PJIIT and Ukrainian technical universities in Kiev and Lviv in the years 2004 – 2006. Its aim was to transfer the curriculum of online undergraduate studies in computer science at PJIIT to the e-education centres established at selected Ukrainian universities and to transfer the System Edu for management of online courses to the established e-centres.

There were the following premises of moving online undergraduate curriculum to Ukraine: IT knowledge is general enough, it requires only translating didactic materials into Ukrainian, universal form of applied technology during development of content makes it easy to move them to another program of studies, modern technology based on Internet and on tele/video conferences is available.

At the beginning we conducted the analysis of Ukrainian education market in order to find institutions which might host e-centers in cooperation with PJIIT. Comparison of curricula of PJIIT and Ukrainian universities was made.

We found that Ukraine experienced need for IT specialists and for the international diploma of completing the studies in IT. Technical universities of Kiev and Lviv were introducing techniques of e-education to their teaching activities.

In February of 2005 the agreements were signed for cooperation between PJIIT and technical universities of Kiev and Lviv and for creating two e-centers at those universities.

A training course on e-learning methods was conducted in the PJIIT building on July 11–15, 2005, with 20 Ukrainian academic teachers taking part (from Kiev, Lviv, and Odessa technical universities).

Purchasing and instalment of hardware and software in the Ukrainian centers took place. The Ukrainian version of the Edu System was prepared and installed in both the Ukrainian e-centers (Fall 2005).

During the years 2005 and 2006 the multimedia materials for online courses were translated from Polish into English and next into Ukrainian language.

In the spring semester of the academic year 2005/2006 two pilot courses of the selected subjects for the selected groups of Ukrainian students were prepared and conducted.

In addition to the UNDP project another project sponsored by the Polish Ministry of Foreign Affairs was also run in the year 2006 whose aim was to include Odessa Polytechnic into the framework of the UNDP project e-centers, to train next group of Ukrainian academic teachers in e-learning methods, and to prepare multimedia materials for 11 humanistic courses which are obligatory in the Ukrainian academic curriculum.

In the year 2007 a similar project sponsored by the Polish Ministry of Foreign Affairs was run whose aim was to include Kharkiv University of Radioelectronics into the framework of the UNDP project e-centres and to train next group of Ukrainian academic teachers in e-learning methods.

It is expected that the Ukrainian e-centers in cooperation with PJIIT will start online studies from the academic year 2008/09.

References

1. Galwas BA (2002) Education over Internet – the end of the beginning. MEW, Technical University of Warsaw, nr 1/2002 (in Polish)
2. McCormack C, Jones D (1998) Building a Web-Based Education System. Wiley Computer Publishing
3. Banachowski L, Mrówka-Matejewska E, Lenkiewicz P (2004) Teaching Information Technology over Internet Proceedings of the conference "Akademia on-line", Bronisławów, May 2004, WSHE Publishing House (in Polish)
4. Transfer of Information Technology to Ukraine. (2004) Project Document United Nations Development Programme
5. Banachowski L, Drabik A, Nowacki JP (2005) About UNDP Project of conducting studies over Internet in Ukraine. V Conference "Virtual University: VU'2005", Technical University of Warsaw (in Polish)

Chapter 15

Hierarchical Bayesian approach for ranking of accident blackspots with reference to cost of accidents

Noorizam Daud[1], Kamarulzaman Ibrahim[2]

[1]Faculty of Information Technology & Quantitative Sciences,
UiTM, Shah Alam, 40450, Selangor, Malaysia, noorzam@tmsk.uitm.edu.my
[2]Statistics Program/Engineering Mathematics Research Group,
Faculty of Science & Technology, UKM, Bangi, 43600, Selangor, Malasia,
kamarulz@pkrisc.cc.ukm.my

Abstract. Road accident is an unfortunate event which is a matter of serious concern to the authority. A proactive measure taken in reducing the rate of accidents is to identify hazardous locations for treatment. In order to allocate resources wisely when treating accident locations, engineers usually rank accident locations based on the mean number of accidents observed over a period of time. Identification, ranking, and selecting hazardous accident locations from a group under consideration is a fundamental goal for traffic safety researchers. The search of a better method to carry out such tasks is the main aim of this study in order to improve road safety in the country. The number of accident varies within and between locations, hence making Bayesian hierarchical model suitable to be applied when allowing for these two stages of variation. This study will illustrate the use of posterior mean to rank accident blackspots.

Keywords. Bayesian hierarchical model, posterior mean, accident blackspots, gamma distribution, Poisson distribution, traffic safety research

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_15,

15.1 Introduction

Since 1990 Ministry of Science, Technology and Environment have been funding research programs to improve the accident data collection and analysis system in Malaysia. The programs also aim to encourage wider usage of the system to assist in the identification of accident blackspots prior to any effective treatment given in order to improve road safety in the country [9]. The current method used in the country to identify such hazardous locations is not based on specific probabilistic approach. Since accidents are random and multi-factor events, the use of probability and tools of statistics in such road safety research is more appropriate. Recently, empirical Bayes methods have been used in road safety studies to identify dangerous locations arguing that adjusting historical data by statistical estimates yields improved predictability [4, 7]. Furthermore, the recent use of ranking procedures based on a hierarchical Bayes approach has been proposed in literature [5, 6, 10], since this method can handle uncertainty and variability in accident data by producing a probabilistic ranking of the accident locations. This chapter will highlight the use of Bayesian hierarchical approach to produce an alternative ranking method in identifying the hazardous accident locations. In addition, it could enable to determine the impact on the ranking when alternative criteria are used. The hierarchical Bayesian method proposed by Schlüter [10] is reviewed and some adjustments have been made by including the fatal and serious injury accident categories and also the ratio of the cost of fatal accidents as compared to serious injury accidents.

15.2 Data

The Royal Malaysian Police has classified accident into four types: fatal, serious injury, slight injury, and damage to the vehicles or properties only [2]. Due to the problem of misclassification of the type of injury accidents, only the accident data for the fatal and serious injury accidents are considered in this study. The analysis made to illustrate the proposed ranking method is based on accident data collected over a 3-year period from 1996 to 1998 for 30 locations. Since the details pertaining to cost of a particular accident may not be readily available, a sample of insurance claims for fatal and serious injury accidents are used for estimating the accident cost and the ratio between these insurance claims will be used as a scaling factor, treated as the nuisance factor in the Bayesian modeling.

15.3 Cost ratio of fatal to serious injury accident

We assume that the insurance claim for fatal accident (C_1) and the insurance claim for serious injury accident (C_2) follow gamma distributions with parameters γ_1, η_1 and γ_2, η_2, respectively.

Consider a new variable $A = \frac{C_1}{C_2}$. By using the method of transformation of variables, the conditional probability of a given $\gamma_1, \gamma_2, \eta_1, \eta_2$ could be written as

$$f(a|\gamma_1,\gamma_2,\eta_1,\eta_2)=\frac{\Gamma(\gamma_1+\gamma_2)\eta_1^{\gamma_1}\eta_2^{\gamma_2}a^{\gamma_1-1}}{\Gamma(\gamma_1)\Gamma(\gamma_2)\ [a\eta_1+\eta_2]^{\gamma_1+\gamma_2}} \qquad ;a>0. \tag{15.1}$$

Based on the maximum likelihood method, the estimates for the parameters $\gamma_1, \gamma_2, \eta_1, \eta_2$ are obtained, where $\hat{\gamma}_1 = 1.0951$, $\hat{\gamma}_2 = 1.1825$, $\hat{\eta}_1 = 12481.8508$, and $\hat{\eta}_2 = 704.1688$. These values are substituted in Eq. (15.1), and by numerical method the median and mean values are found to be 1.31 and 8.6, respectively [8].

15.4 The hierarchical Bayesian model

Consider two discrete random variables X_{ij} and Y_{ij}, each representing the number of fatal accidents and the number of non-fatal accidents occurring at locations $i = 1,2,\ldots,k$ in $j = 1,2,\ldots,t_i$ years. Since each location observed two accident categories, random variables X_{ij} and Y_{ij} are assumed to have a mean number of accident per year of λ_{1i} and $\lambda_{2i2,}$ respectively. Since both X_{ij} and Y_{ij} satisfy the characteristics of a Poisson process, it is reasonable to assume that both variables are following Poisson distribution with mean λ_{1i} and λ_{2i}, respectively.

Let the number of fatal accidents occurring in location i in the period t_i year be denoted as $X_i = \sum_{j=1}^{t_i} X_{ij}$, and the number of serious injury accidents occurring at location i in t_i year be denoted as $Y_i = \sum_{j=1}^{t_i} Y_{ij}$. Hence, random variable X_i and $Y_{i,}$ respectively, are assumed to be having Poisson distribution with mean number of accidents of $t_i\lambda_{1i}$ and $t_i\lambda_{2i}$ given as follows:

$$f(x_i \mid \lambda_{1i}) = \frac{(t_i\,\lambda_{1i})^{x_i}\ \exp(-t_i\lambda_{1i})}{x_i!}, \quad x_i = 0,1,2.... \tag{15.2}$$

and $$f(y_i \mid \lambda_{2i}) = \frac{(t_i\,\lambda_{2i})^{y_i}\ \exp(-t_i\lambda_{2i})}{y_i!}, \quad y_i = 0,1,2,...\ ,$$

for i=1,2,…,k where $\lambda_{1i} > 0$ and $\lambda_{2i} > 0$.

In the following explanation to obtain the mean posterior, t will be excluded since it is regarded as a constant term.

Assume that the uncertainties in the mean number of fatal and serious injury accidents are modeled as gamma distributions, which are also commonly known as the conjugate priors.

Based on the elicitation of expert opinions and referring to some previous studies [1, 3, 5], it reveals that the expected cost for fatal accident is more than the expected cost of serious injury accident. Since the true cost of each type of accident is not known, we consider that the ratio of claims for fatal accident and serious injury accident as obtained in Eq. (15.1) could be used as a scaling factor for scaling up the expected number of fatal accidents, thus adjusting the hazardous level of each location.

The joint posterior distribution of λ_{1i}, λ_{2i} and a conditional on X_i, Y_i could be obtained through the Bayes theorem mechanism given as

$$f(\lambda_{1i},\lambda_{2i},a \mid X_i,Y_i) \propto f(X_i,Y_i \mid a,\lambda_{1i},\lambda_{2i})\, f(a,\lambda_{1i},\lambda_{2i}) \tag{15.3}$$

Since the parameters λ_{1i}, λ_{2i}, and a are assumed independent, Eq. (15.3) could be simplified as follows:

$$f(\lambda_{1i},\lambda_{2i},a \mid X_i,Y_i) \tag{15.4}$$
$$\propto f(X_i \mid a,\lambda_{1i})\, f(Y_i \mid \lambda_{2i})\, f(\lambda_{1i} \mid \alpha_1,\beta_1)\, f(\lambda_{2i} \mid \alpha_2,\beta_2)\, f(a).$$

Let $\lambda_i' = a\lambda_{1i} + \lambda_{2i}$ represent the prioritized score to be used in ranking the accident locations. Hence, a will be regarded as a nuisance factor and it should be integrated out.

The posterior mean of λ_i' which is the required prioritized score could be obtained as

$$E(\lambda' \mid X_i,Y) = E(a \mid X_i,Y_i)\ E(\lambda_{1i} \mid X_i,Y_i) + E(\lambda_{2i} \mid X_i,Y_i) \tag{15.5}$$

On the other hand, if the factor a is regarded as a constant, then Eq. (15.5) will be further reduced as follows:

$$E(\lambda_i' \mid X_i, Y_i) = aE(\lambda_{1i} \mid X_i, Y_i) + E(\lambda_{2i} \mid X_i, Y_i) \tag{15.6}$$

For comparison on the sensitivity of the results to the choice of the prior distributions, four different prior distributions are considered. The prior distributions are

(i) Prior 1: Ratio of two gamma distributions (see Eq. (15.1)).
(ii) Prior 2: Improper prior.
(iii) Prior 3: Median value (1.31).
(iv) Prior 4: Mean value (8.6).

15.5 Discussion of results

From Table 15.1, it appears that the results slightly change when different prior distributions are used. As expected, the results based on the choice of improper prior are similar to those obtained based on not allowing for factor *a* in the model. When factor *a* is considered as a constant, with allowance of the median value of *A*=1.31 and mean of *A*=8.6, respectively, it is found that the uncertainty of the estimated posterior mean is much larger in the case when the later prior is used. Thus, when the two measures of *A* are compared, median is a better choice. On the average, it is found that the posterior standard deviation for the estimated posterior mean that is based on prior 1 is smaller compared to when other priors are used. We believe that allowance for cost of accident is a prudent way of ranking of blackspots.

15.6 Conclusions

As mentioned by Schlüter [10], the ranking based on the posterior mean will differ from the ranking based on observed rates; hence, making the hierarchical Bayesian approach more appealing since it is more flexible when the variations in the data are taken into account. The results show that the cost of accidents should be considered in the selection and ranking of hazardous accident locations. The ranking using posterior mean values can provide policy makers with a scientific instrument which is statistically sound to select hazardous road locations. By allowance for cost of accident in addition to the number of accidents, accident locations can be ranked more precisely. It is hoped that this proposed method will assist the authority in identifying the blackspots which require prompt attention.

Table 15.1. Posterior mean $E(\lambda' \mid X_i, Y_i, a)$ for 30 locations using several priors for factor *a*

Location	X	Y	Prior 1	Prior 2	Prior 3	Prior 4	Without factor a
L[1]	8	12	4.808	5.383	5.679	7.491	5.367
L[2]	4	19	6.135	6.625	6.837	8.107	6.598
L[3]	6	17	5.883	6.379	6.621	8.193	6.354
L[4]	11	10	4.579	5.259	5.627	7.819	5.262
L[5]	3	16	5.307	5.766	5.968	7.066	5.766
L[6]	3	17	5.555	5.996	6.195	7.334	6.021
L[7]	8	13	5.09	5.642	5.927	7.761	5.639
L[8]	4	19	6.164	6.638	6.818	8.119	6.62
L[9]	9	12	4.907	5.521	5.807	7.799	5.513
L[10]	8	10	4.315	4.895	5.18	6.994	4.904
L[11]	3	15	5.082	5.522	5.698	6.834	5.518
L[12]	3	17	5.531	5.999	6.222	7.36	6.005
L[13]	3	19	6.041	6.515	6.656	7.831	6.493
L[14]	4	12	4.442	4.9	5.122	6.401	4.9
L[15]	8	10	4.33	4.916	5.192	6.987	4.913
L[16]	5	18	6.001	6.511	6.767	8.131	6.494
L[17]	2	19	5.899	6.376	6.528	7.549	6.388
L[18]	3	14	4.809	5.246	5.487	6.642	5.271
L[19]	2	18	5.656	6.125	6.298	7.341	6.145
L[20]	2	18	5.686	6.137	6.323	7.29	6.131
L[21]	5	9	3.801	4.296	4.546	5.973	4.312
L[22]	0	20	5.827	6.374	6.496	7.261	6.381
L[23]	4	12	4.451	4.919	5.111	6.415	4.917
L[24]	2	16	5.188	5.633	5.822	6.819	5.648
L[25]	2	17	5.421	5.869	6.068	7.08	5.881
L[26]	3	13	4.566	5.022	5.235	6.359	5.052
L[27]	8	5	3.095	3.681	3.973	5.777	3.684
L[28]	1	17	5.268	5.768	5.924	6.794	5.777
L[29]	3	15	5.057	5.517	5.72	6.824	5.518
L[30]	4	13	4.808	5.152	5.368	7.491	5.177

Prior 1: Ratio of two gamma distributions
Prior 2: Improper Prior
Prior 3: Median value (a = 1.31)
Prior 4: Mean value (a = 8.6)

Acknowledgement. The author would like to thank the Malaysian Road Ministry of Works and the Malaysian Royal Police for providing the accident data to be used in the study. Special thanks also go to the Engineering Mathematics Research Group of UKM and UiTM.

References

1. Al-Masaeid HR, Al-Mashakbeh AA, Qudah AM (1999) Economic costs of traffic accidents in Jordan. Accident Analysis and Prevention 31:347–357
2. Baguley CJ (1995) Interim Guide on Identifying, Prioritising and Treating Hazardous Locations on Roads in Malaysia. Public Works Department, Malaysia
3. Downing A (1997) Accident cost in Indonesia: A review. Overseas Centre Transport Research Laboratory Crowthorne Berkshire United Kingdom, RG456AU
4. Elvik R (1997) Evaluations of road accident blackspot treatment: A case of the iron law of evaluation studies? Accident Analysis and Prevention 29(2):19–199
5. Geurts K, Wets G, Vanhoof K (2005) Ranking and selecting dangerous accident locations: Case study. Urban Transport 77:229–238
6. MacNab YC (2003) A Bayesian hierarchical model for accident and injury surveillance. Accident Analysis and Prevention 35:91–102
7. Miaou SP (1994) The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accident Analysis and Prevention 26(4):471–482
8. Noorizam Daud, Kamarulzaman Ibrahim (2007) Taburan Nisbah Kos Kemalangan Maut dan Cedera Parah. Jurnal Teknologi Maklumat dan Sains Kuantitatif. Universiti Teknologi MARA. Jilid 6
9. Radin Umar (1998) Critical review of the status of road safety in Malaysia. Journal Chartered Institute of Transport, UK 7(1):20–40
10. Schlüter PJ, Deely JJ, Nicholson J (1997) Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. The Statistician 46(3):293–316
11. Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2003) WinBUGS Version 1.4 User Manual. Cambridge: Medical Research Council & Imperial College, UK. http://www.mrc-bsu.cam.ac.uk/bugs

Part 2

Circuits, systems, electronics, control and signal processing

Neural networks represent a powerful data processing technique that has reached ripeness and broad application. When clearly understood and appropriately used, they are a compulsory component for the best use of the available data, in order to build models, make predictions, process signals and images, recognize shapes, etc. This part is written by experts in the application of the neural networks in different areas. The chapters in this part cover the problems of the application of the artificial neural networks like handwriting knowledge based on parameterization for writer identification, identification surfaces family, neuro-fuzzy models and tobacco control, PNN for molecular level selection detection, automatically fine-tune artificial neural networks parameters, a neuro-fuzzy network for supporting detection of diabetic symptoms, and empirical assessment of LR- and ANN-based fault prediction techniques.

Chapter 16

Impulsive noise removal image ehancement technique

Subrajeet Mohapatra, Pankaj Kumar Sa, Banshidhar Majhi

CSE, NIT Rourkela – 769008, India
subrajeets@gmail.com, pankajksa@nitrkl.ac.in, a, bmajhi@nitrkl.ac.in

Abstract. Several areas like remote sensing, biomedical analysis, and computer vision require good image contrast and details for better interpretations and diagnoses's. In the literature various image enhancement algorithms have been proposed to improve the perceptual aspects of the image for poorly contrasted images. The perceptual appearance of an image may be significantly improved by modifying the high-frequency components to have better edge and detail information in the image. The proposed scheme is a modification of simple unsharp mask image enhancement technique. Comparative analysis on standard images at different noise conditions shows that the proposed scheme, in general, outperforms the existing schemes.

Keywords. Image enhancement, Contrast enhancement, Unsharp masking, Impulsive noise removal

16.1 Introduction

Image enhancement algorithms are used to increase the visibility of images for their specific applications. A good number of methods are available in the literature for enhancing different properties or components of images [1, 2]. Contrast enhancement techniques can be classified into intensity-based techniques and feature-based techniques. Intensity-based techniques can be modeled using the relation (see Eq. 16.1)

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_16,

$$I'(x,y) = f(I(x,y)) \tag{16.1}$$

where $I(x,y)$ is the original image and $I'(x,y)$ is the output image after enhancement and f is the transformation function applied to the whole image. Contrast stretching [2], histogram equalization [3] are the popular members within this category. Histogram equalization is a widely accepted image enhancement technique. Variations of this algorithm like bihistogram equalization [4], multipeak histogram equalization [5], adaptive histogram equalization are also available in the literature. Adaptive histogram equalization (AHE) [6] is the widely accepted version of histogram equalization which overcomes the pitfalls of general histogram equalization. The objective of feature-based enhancement technique is to enhance the high-frequency component or the image details of a poor-contrast image. Feature-based methods can be expressed using the relation

$$I'(x,y) = L(x,y) + \lambda.H(x,y) \tag{16.2}$$

where L and H represent the low-frequency and high-frequency components, respectively. λ is the enhancement gain or amplification factor required for amplifying the high-frequency components of the image for better perception.

16.1.1 Linear unsharp masking

Linear unsharp masking (UM) [1] is an important scheme in the feature-based image enhancement category. In the UM technique, a high-pass filtered scaled version of the original input image is added to itself as shown in Fig. 16.1 to obtain an enhanced version.

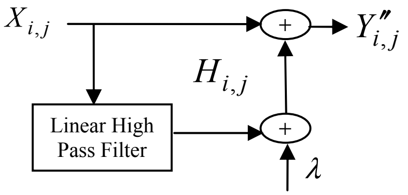


Fig. 16.1. Simple linear unsharp masking

Although this simple method works well in many applications it suffers a major drawback which limits its performance in certain applications. There is every possibility that the noise is also amplified along with the edge and detail features of the image since we are using a scaled version of the high-frequency components of an image. So the objective is to remove impulsive noise followed by applying unsharp masking while preserving edge details. In our proposed scheme we use a FLANN to determine impulse noise threshold so that we can apply selective filtering on the noisy pixels only and prevent unnecessary loss of image details.

In what follows the detail of the proposed scheme along with the algorithm is presented in Sect. 16.2. Adaptive threshold selection using CV is discussed in Sect. 16.3. Finally, Sect. 16.4 presents the simulation results and Sect. 16.5 gives the concluding remarks.

16.2 Proposed scheme

The proposed scheme consists of impulse detection and iterative filtration followed by image contrast enhancement as shown in Fig. 16.2.

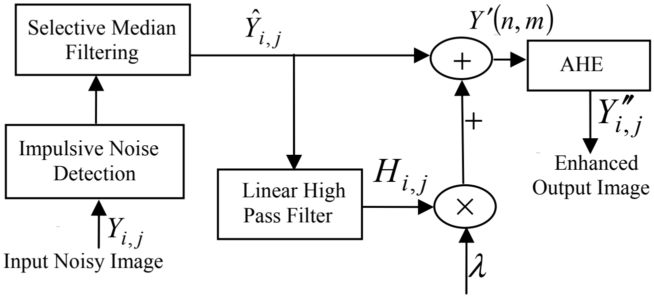


Fig. 16.2. Proposed model for image contrast enhancement

16.2.1 Impulse detection

Images are frequently contaminated by impulsive noise due to noisy sensors or channel transmission errors [1,2]. Since use of high-pass filter in unsharp masking (UM) [1] the scheme becomes highly sensitive to noise. There are many types of impulsive noise. Let $X_{i,j}$ be the gray level of an original image X at pixel location (i,j) and $[n_{\min}, n_{\max}]$ be the dynamic range of X. Let $Y_{i,j}$ be the gray level of the noisy image Y at pixel (i,j), and then the random-valued impulsive noise may be defined as

$$Y_{i,j} = \begin{cases} X_{i,j} & \text{with probability } 1-p \\ R_{i,j} & \text{with probability } p \end{cases} \tag{16.3}$$

where $R_{i,j} \in [n_{\min}, n_{\max}]$ and p is the noise ratio. Whereas for *fixed-valued impulsive noise* (better known as *salt and pepper noise*) $R_{i,j} \in \{n_{\min}, n_{\max}\}$. It is usually seen that removal of *salt and pepper noise* is easier in comparison to RVIN since behavior of RVIN pixels and its surrounding pixels is very similar. In this chapter we focus only on *random-valued impulsive noise* where $Y_{i,j}$ can be of any value from $n_{\min}$ to $n_{\max}$.

In the proposed scheme an impulsive noise detector based on second-order difference is used to determine the threshold for impulse noise detection. Median filtration is performed selectively based on the decision of the threshold. The following mathematical formulation describes whether to filter or to skip a pixel located at (i, j) of a test window:

$$\hat{Y}_{i,j} = \begin{cases} Y_{i,j} & d_{i,j} = 1 \\ Z_{i,j} & d_{i,j} = 0 \end{cases} \tag{16.4}$$

where $Z_{i,j} = \text{median } \{Y_{i-k,j-l}, (k,l) \in W\}$ and W is a predetermined window, usually of size 3×5 or 5×5 [2]. The filtration is performed selectively based on the decision index $d_{i,j}$ which controls the filtering operation.

Algorithm

Pass one

1. Choose a test window $Y^{(T)}$ of size 3×5 centered at (i, j) of Y Choose a sub window $Y^{(W)}$ of size 3×3 centered at (i, j) of $Y^{(T)}$.
2. Compute the first-order 3×4 difference matrix fd.

$$fd_{i+k,j+l} = Y^{(T)}_{i+k,j+l} - Y^{(T)}_{i+k,j+l-1} \tag{16.5}$$

   where $k = -1, 0, 1$ and $k = -1, 0, 1, 2$
3. Compute the second-order 3×3 difference matrix sd from fd.

$$sd_{i+r,j+s} = fd_{i+r,j+s+1} - fd_{i+r,j+s} \tag{16.6}$$

   where $r = -1, 0, 1$ and $s = -1, 0, 1$

4. Compute the decision parameter d

$$d_{i,j} = \begin{cases} 0 & \text{if } \left|sd_{i,j}\right| > \theta_1 \\ 1 & \text{otherwise} \end{cases} \tag{16.7}$$

If $d_{i,j}$ is zero, replace the $Y_{i,j}$ pixel with the median value of its neighborhood, otherwise leave it as it is.

5. Repeat the above steps for each window from top-left to bottom-right corner of the noisy image.

Pass two

The window $Y^{(T)}$ selected is of size 5×3 centered at (i, j) of Y and sub window $Y^{(W)}$ of size 3×3 centered at (i, j) of $Y^{(T)}$. The first-and second-order differences are calculated in vertical fashion and the decision index is determined, followed by selective filtration similar to the steps described earlier. The threshold values taken here in this pass is θ_2. The threshold values θ_1 and θ_2 are obtained using FLANN as described in Sect. 16.3.

All the steps in the second iteration are repeated for each test window column wise from top-left to bottom-right corner of the image obtained from pass one. Then we perform image enhancement as described in Sect. 16.2.2.

16.2.2 Image contrast enhancement

The filtered image output $\hat{Y}_{i,j}$ is fed into a high-pass filter to separate the high-and low-frequency components. We can choose a suitable gain factor for amplification of image detail regions like sharp edges depending on the application. So this can be represented using the relation given as follows:

$$Y'_{i,j} = \hat{Y}_{i,j} + \lambda H_{i,j} \tag{16.8}$$

where $H_{i,j}$ is the output of a linear high-pass filter (see Eq. 16.9).

$$H_{i,j} = 4\hat{Y}_{i,j} - \hat{Y}_{i-1,j} - \hat{Y}_{i+1,j} - \hat{Y}_{i,j-1} - \hat{Y}_{i,j+1} \tag{16.9}$$

λ is the positive gain factor that controls the level of enhancement required by an application, this may vary from one application to other. Using the proposed filtering scheme we are able to preserve the image edge and detailed features. Adaptive histogram equalization (AHE) technique is further applied on $Y'_{i,j}$ for better visual perception to obtain the desired image $Y''_{i,j}$.

16.3 Adaptive threshold selection

Artificial neural networks (ANN) has emerged as a powerful learning technique to perform complex tasks in highly nonlinear dynamic environments. Once trained under supervision, the ANN has the capability to generalize and predict the output for any given input in similar type of problems [7]. Numerous structural variations of ANN are available in the literature [8]. A variation of ANN is the functional linked artificial neural network (FLANN), which is a flat net without any hidden layers [8, 9]. The advantage of using a reduced neural network like FLANN is less costly and faster in operation. Training of FLANN by BPA is very simple having lesser computational load and faster convergence rate. The functional expansion increases the dimension of the input vector that in turn improves discrimination capability of the hyper planes generated by the FLANN.

The proposed impulsive noise detector is shown in Fig. 16.1, which is a two-layered structure. The input to the network is a global coefficient of variance (CV) [1] of the noisy image calculated using the relation (see Eq. 16.10)

$$\mathrm{CV} = \sigma / \mu \tag{16.10}$$

σ, μ are the global standard deviation and mean, respectively, of the image. The input CV is functionally expanded in the input layer with the trigonometric polynomial basis function (see Eq. 16.11).

$$\begin{gathered} 1, \sin(\pi CV), \sin(2\pi CV), \ldots, \sin(N\pi CV), \\ CV, \cos(\pi CV), \cos(2\pi CV), \ldots, \cos(N\pi CV) \end{gathered} \tag{16.11}$$

In order to determine the error we compare the actual output of the network with the desired output. As per the error value we update the weight matrix between input and output layers using back propagation algorithm.

We take an image say pepper that is corrupted with impulsive noise of noise density between 0.01 and 0.30 in steps of 0.05. Each corrupted image is subjected to the proposed filter varying the threshold from 0 to 1 in steps of 0.01 and the corresponding mean squared error (MSE) value is computed using Eq. (16.12).

$$\mathrm{MSE} = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y) - \hat{f}(x,y)\right) \tag{16.12}$$

where $(M \times N)$ is the size of the image $X_{i,j}$ and $Y_{i,j}$, which represents the pixel values at $(i, j)_{th}$ location of original image and restored image, respectively. The minimum MSE and the corresponding threshold value called optimum threshold $(\theta_{optimum})$ are recorded. Since the MSE requires the original image for computation we cannot use it for threshold detection in real-time applications. Here we use CV in place of mean and variance which can be easily computed from the noisy image available and is used for threshold prediction using a FLANN (Fig. 16.3).

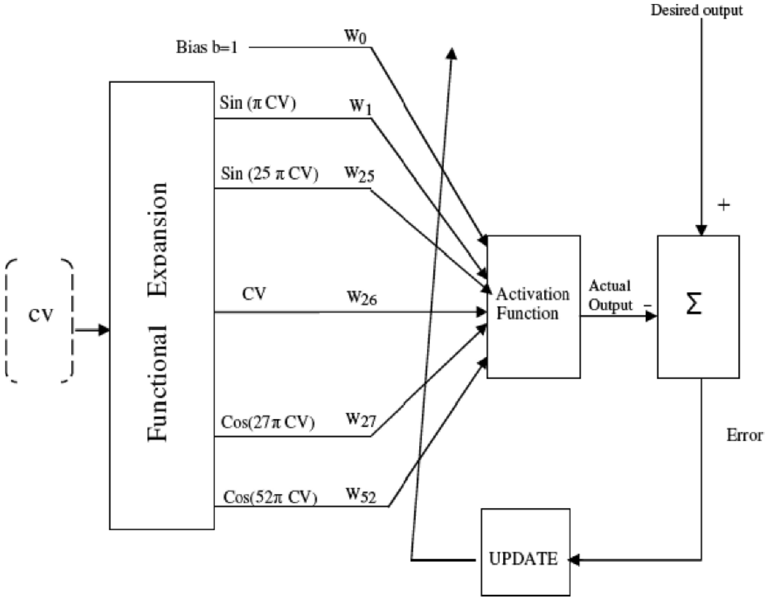


Fig. 16.3. FLANN structure for adaptive threshold selection

For training the FLANN the input–output patterns $((CV) \rightarrow \theta_{Optimum})$ for different noise levels are generated for different images like *Lena, Lisa, Boat*, etc. The training convergence for functional link artificial neural network (FLANN) is shown in Fig. 16.4.

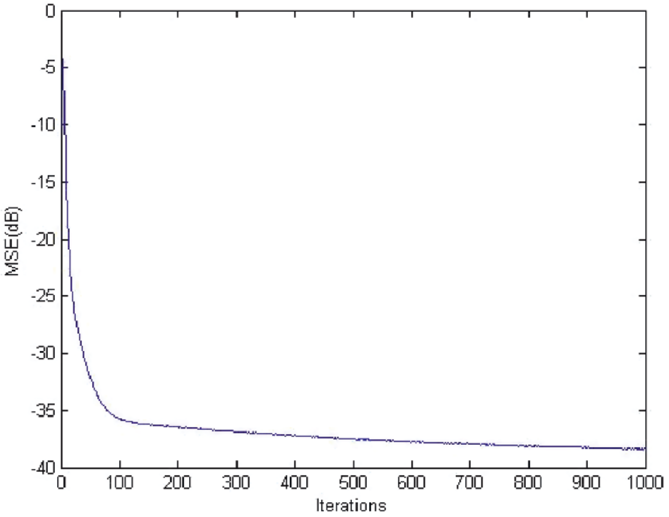


Fig. 16.4. Convergence characteristics of FLANN

16.4 Simulation results

The superiority of the proposed scheme is demonstrated by conducting two experiments. Peak signal-to-noise ratio (PSNR) in dB, as defined in Eq. (16.13) is the metric used to compare the noise removal capability of the proposed scheme with the existing schemes.

$$\mathrm{PSNR} = 10\log_{10}\left(255^2/\mathrm{MSE}\right) dB \tag{16.13}$$

where MSE is the mean squared error as defined in Eq. (16.12). Subjective results for Lena, Boat, and Pepper are shown for comparing image enhancement procedure.

16.4.1 Experiment I

Lena image is corrupted with noise ranging from 0.01 probability to 0.30. Various standard schemes like progressive switching median(PSM) [10], adaptive center weighted median filter (ACWMF) [11], two-pass (2-pass) [12], switching median (SWM(5×5)), accurate noise detector (AND) [13], two-output nonlinear filter (2-OUTPUT) [14], median rational hybrid filter-II (MRHF2) [15], detail preserving impulsive noise removal

(DPINR) [16], median rational hybrid filter-II (MRHF2) [17], impulse detection based on pixel-wise MAD(PWMAD) [16], FLANN-based adaptive threshold selection for detection of impulsive noise in Images (FLANN-ATS) [18] are simulated along with the proposed scheme. PSNR obtained from various schemes for Lena image are plotted and shown in Fig. 16.5. Table 16.1 depicts the comparative study of PSNR values for standard images, viz. *Lena, Lisa, Boat,* and *Clown,* that the performance of the proposed noise removal scheme is superior to existing schemes.

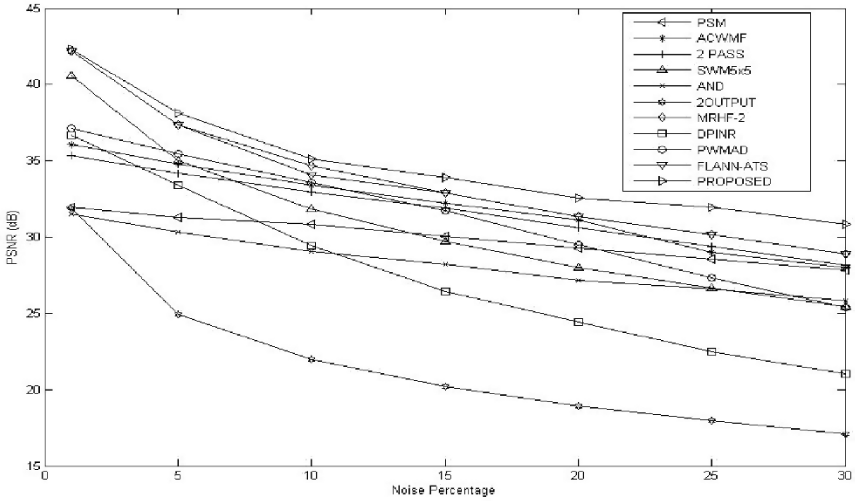


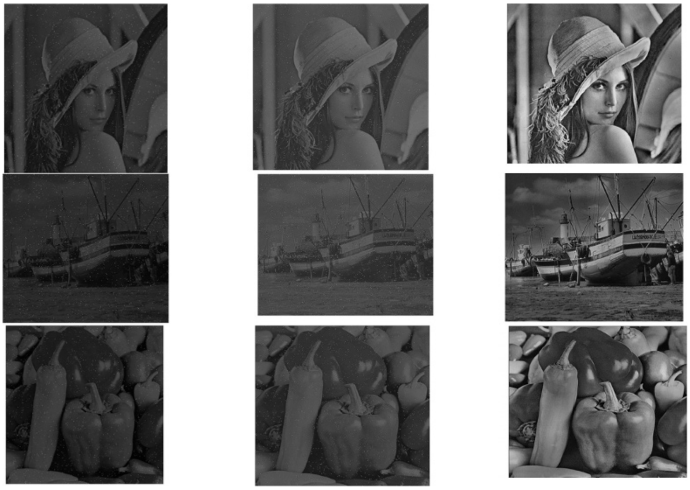
Fig. 16.5. PSNR obtained from various schemes for Lena image

Table 16.1. Comparative results in PSNR (dB) of filtering images corrupted with 15% of noise

Images	PSM	ACWMF	2-Pass	SWM	AND	2-OUTPUT	MRHF2	DPINR	PWMAD	FLANN -ATS	Proposed
Lena	30.05	32.21	31.88	29.71	28.23	20.21	25.47	26.62	31.73	32.89	33.89
Lisa	29.09	31.78	31.34	28.31	26.98	20.95	26.90	27.38	30.50	30.84	31 .63
Boat	26.78	28.87	29.33	28.34	25.19	20.27	25.14	25.87	29.25	29.93	30.49
Clown	21.70	22.56	22.84	24.06	21.31	17.76	19.24	20.99	23.02	24.24	25.13

16.4.2 Experiment II

To visualize the subjective image enhancement performance, the enhanced Lena, Boat, Pepper images are compared with the results of the simple linear unsharp masking [1] and is shown in Fig.16.6. Since we do not have any quantitative evaluation measure for image enhancement because of absence of any ideal image we are forced to go for subjective evaluation. And it can be easily realized that since we are not amplifying the noise and also preserving the image details while filtering the proposed method is much better in comparison to simple unsharp masking. Further we apply AHE for better visual perception. So the proposed scheme gives better performance in comparison to simple unsharp masking.



(a) Low contrast (b) Linear unsharp masking (c) Proposed

Fig. 16.6. Subjective comparison of enhanced images for *Lena, Boat*, and *Pepper*

16.5 Conclusions

This chapter has proposed a novel filtering scheme for suppressing impulsive noise from contaminated images along with provision for better image contrast. Since we are using selective median filtering this scheme is able to preserve the image details for further image enhancement after impulse noise removal. Through exhaustive computer simulations it is observed that the proposed scheme exhibits superior performance over other schemes.

References

1. Jain AK (1989) Fundamentals of Image Processing. Prentice Hall, 2003
2. Gonzalez RC, Woods RE (1992) Digital Image Processing. Addison Wesley
3. Kim YT (1997) Contrast enhancement using brightness preserving bi-histogram equalization. IEEE Transactions on Consumer Electronics 43(1):1–8
4. Wongsritong K, Kittayaruasiriwat K, Cheevasuvit F, Dejhan K, Somboonkaew A (1998) Contrast enhancement using multipeak histogram equalization with brightness preserving. Circuits and Systems, IEEE APCCAS
5. Brownrigg DRK (1984) The weighted median filter. Comm. ACM 27:807–818
6. Lee JS (1980) Digital image enhancement and noise filtering by using local statistics. IEEE Transactions on Pattern Analysis Machine Intelligence PAMI:165–168
7. Pizer SM, Amburn EP (1987) Adaptive histogram equalization and its variations. Computer Vision, Grpahics, and Image Processing 39:355–368
8. Patra JC, Panda G, Baliarsingh R (1994) Artificial neural network based nonlinearity estimation of pressure sensors. IEEE Trans Instrum Meas 3:874–81
9. Majhi B, Sa PK (2006) FLANN-based adaptive threshold selection for detection of impulsive noise in images. AEU – Inernational Journal of Electronics and Communication 61:478–484
10. Xu X, Miller EL (2002) Adaptive two-pass median filter to remove impulsive noise. Proceedings of International Conference on Image Processing 1:22–25
11. Chen T, Wu HR (2001) Adaptive impulse detection using center weighted median filters. IEEE Signal Process Letters 8:1–3
12. Kondo K, Haseyama M, Kitajima H (2002) An accurate noise detector for image restoration. Proceedings of International Conference in Image Processing I-321–I-324
13. Russo F (2004) Impulse noise cancellation in image data using a two-output nonlinear filters. Measurement 36:205–213
14. Khriji L, Gabbouj M (1998) Median-rational hybrid filters. Proceedings of International Conference on Image Processing 2:853–857
15. Crnojevic V, Senk V, Trpovski Z (2004) Advanced impulse detection based on pixel-wise MAD. IEEE Signal Processing Letters 11:589–592
16. Zhang S, Karim Md (2002) A new impulse detector for switching median filters. IEEE Signal Processing Letters 9:360–363
17. Alajlan N, Kamel M, Jernigan E (2004) Detail preserving impulsive noise removal. Signal Processing: Image Communication 19:993–1003
18. Wang Z, Zhang D (1999) Progressive switching median filter for the removal of impulse noise from highly corrupted images. IEEE Transactions on Circuits Systems–II: Analog Digital Signal Process 46:78–80

19. Frost VS, Stiles JA, Sharmugan KS, Holtzman JC (1982) A model for radar images and its application to adaptive digital filtering of multiplicative noise. IEEE Transactions on Pattern Analysis Machine Intelligence PAMI-4:157–165
20. Haykin S (2003) Neural Networks (2nd ed.). Prentice-Hall
21. Narendra PM, Fitch RC (1981) Real-time adaptive contrast enhancement. IEEE Transactions on Pattern Analysis Machine Intelligence PAMI(3):655–661
22. Patra JC, Pal RN, Chatterji BN, Panda G (1999) Identification of nonlinear dynamic systems using functional link artificial neural networks. IEEE Transactions on Systems Man Cybernet 9:254–262

Chapter 17

Similarity-based model for transliteration

Mohamed Abdel Fattah[1,2], Fuji Ren[1,3]

[1] Faculty of Engineering, University of Tokushima, 2-1 Minamijosanjima Tokushima, 770-8506, Japan, (mohafi, ren)@is.tokushima-u.ac.jp
[2] FIE, Helwan University, Cairo, Egypt
[3] School of Information Engineering, Beijing University of Posts & Telecommunications, Beijing, 100088, China

Abstract. A significant proportion of out of vocabulary (OOV) words are named entities and technical terms. Typical analyses find around 50% of OOV words to be named entities. Yet these can be the most important words in the queries. For example, in the list of queries for TREC 2001 cross-language track, all 25 queries contained proper names. Cross-language retrieval performance (average precision) reduced more than 50% when named entities in the queries were not translated. One way to deal with OOV words when the two languages have different alphabets is to transliterate the unknown words, that is, to render them in the orthography of the second language. Transliteration is the process of formulating a representation of words in one language using the alphabet of another language. In the present study, we present different approaches for transliteration of proper noun pair's extraction from parallel corpora based on different similarity measures between the English and the romanized Arabic proper nouns under consideration. The strength of our new system is that it works well for low-frequency proper noun pairs. We evaluate the presented new approaches using two different English–Arabic parallel corpora. Most of our results outperform previously published results in terms of precision, recall, and F-Measure.

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_17,

17.1 Introduction

Recently, much research has been done on machine transliteration for many language pairs, such as English/Arabic [1,2], English/Chinese [3], English/Japanese [4], and English/Korean [5]. Most of the above approaches require a pronunciation dictionary for converting a source word into a sequence of pronunciations. However, words with unknown pronunciations may cause problems for transliteration. On the other hand, much research has focused on the study of automatic bilingual lexicon construction based on bilingual corpora. Proper nouns and corresponding transliterations can often be found in parallel corpora or topic-related bilingual comparable corpora. However, many methods dealt with this problem based on the frequencies of words appearing in corpora, an approach which cannot be effectively applied to low-frequency words, such as transliterated words [6]. Fung used different approaches to create translation pairs from parallel and comparable corpora [7–9]. For instance, in [7], she presented a pattern matching method for compiling a bilingual lexicon of nouns and proper nouns from unaligned, noisy parallel texts of Asian/Indo-European language pairs. Although the simplicity of the used approach the recall was very small. On the other hand, Fattah et al. [6] presented two algorithms and their combination to automatically extract an English/Arabic bilingual dictionary from parallel texts that exist in the Internet archive after using an Arabic light stemmer as a preprocessing step. Both Fung and Fattah approaches do not require pronunciation dictionary for converting a source word into a sequence of pronunciations and they give reasonable results. Therefore, we have exploited the pattern matching method of Fung [7] and Fattah's approach to extract transliteration pairs from English–Arabic parallel corpus and we used them as baseline methods.

17.1.1 Pattern matching approach

In pattern matching approach, tagging information of one language is used. Word frequency and position information for high- and low-frequency words are represented in two different vector forms for pattern matching.

17.1.2 Combination of algorithms 1 and 2 by Fattah et al. [6]

The first algorithm of Fattah et al. uses a similarity metric S(a, e) between words in Arabic language (A) and words in English language (E) based on statistical co-occurrence and the frequency of each Arabic and English word. Then, it computes the association scores for a set of translation pairs (a, e) ∈ (A, E). Depending on a certain threshold, the translation pairs whose association score exceeds this threshold become the entries in the translation lexicon. The second algorithm of Fattah et al. is based on statistical co-occurrence and the frequency of each Arabic and English word too. However, it can extract translation pairs from two sentence pairs only. This algorithm can capture dependencies between groups of words to get word/phrase translation pair which was the problem of many statistical approaches like the first algorithm. Using the first algorithm, we can achieve high precision with low recall. However, it is difficult to handle the translation of compound nouns. The second algorithm does not have the disadvantages of the first algorithm since it can handle the translation of compound nouns. Moreover the precision and recall are higher than that of the first algorithm. However, the processing time required for the second algorithm is higher than that of the first one. This led us to use a certain combination of algorithm 1 and algorithm 2 to gain the advantages of both of them and avoid the disadvantages as much as possible.

17.2 The proposed English–Arabic proper noun transliteration pairs creation approach

The proposed English–Arabic proper noun transliteration pair's creation system extracts all proper nouns from the English sentence using the CLAWS4 POS tagger http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html. It also extracts all proper nouns from the associated Arabic sentence using the Buckwalter Arabic Morphological Analyzer Version 1.0. All the Arabic proper nouns are romanized using the table in http://archimedes.fas.harvard.edu/mdh/arabic/arabic-loc.pdf. The similarity (based on different similarity measures as will be illustrated in the coming sections) between every English and romanized Arabic proper noun pair is measured. The English–Arabic proper noun pair which has similarity score above a certain threshold (th) is extracted. The system repeats this step for all English and Arabic proper nouns that exist in the sentence pair. The system applies the previous steps on all remaining sentence pairs to create all possible transliteration pairs available in the corpus under consideration. The following pseudo-code illustrates the previously mentioned methodology steps.

The Methodology Pseudo-Code

```
Set ie = ia = n = 1
```

```
E: Extract proper nouns of English_Sentence(n) and Arabic_Sentence(n)
R: Romanize Arabic_Proper_Noun(ia)
Score = SIM (Romanized_Arabic_Proper_Noun(ia), English_Proper_Noun(ie))
      If Score >= th
            Copy Arabic_Proper_Noun(ia) & English_Proper_Noun(ie)with the score value  in a file
      End
            ia= ia +1
      if ia <= na
      GOTO R
      End
Set ia = 1 & ie = ie + 1
      if ie <= ne
      GOTO R
      End
Set ia = ie = 1 & n = n + 1
      If n <= N
      GOTO E
      End
```

Here, *na* and *ne* are the total number of Arabic and English proper nouns in the Arabic and English sentence number *n,* respectively. “th” is a predefined threshold. *SIM* is the similarity measure that will be defined in the following sections. *N* is the total number of English–Arabic sentence pairs. The Arabic proper noun consonant and long vowel letters are romanized according to the table in http://archimedes.fas.harvard.edu/mdh/ arabic/ arabic-loc.pdf.

The following sections describe Dice’s similarity coefficient besides two proposed different similarity measures to measure the similarity between English and romanized Arabic proper nouns.

17.2.1 Dice’s similarity coefficient

Dice’s similarity coefficient was originally developed in the field of biology to describe the degree of similarity between two species of plants according to the number of features (such as hairy stems) that they had in common. McEnery and Oakes [10] used Dice’s similarity coefficient to describe the degree of similarity between a word in one language and its translation in another.

17.2.2 Similarity measure 1 (SIM1)

Most Arabic words have a syllable of CV. Most of the Arabic words contain a short or long vowel between two consonant letters. Take the Arabic word "محمد" "mohammad" as an example. The short vowels 'o', 'a' and 'a' existed between the consonants "m, h," "h, m", and "m, d," respectively. Moreover the short vowels do not appear on the Arabic words in almost all Arabic documents. Hence, Dice's approach to measure similarity between English–Arabic transliteration pairs does not work well. We have decided to use our proposed similarity measure called "SIM1." Using SIM1, the system specifies the similarity score between the English and the romanized Arabic words by matching the consonant characters of the English word with the romanized Arabic characters. The system excludes the effect of vowel between two successive consonants and the repeated consonants by using the following algorithm to specify SIM1:

```
Set SIM1 = 0
Set ia = ie = 0
R:    Read the Romanized Arabic character(ia)
      Read the English character(ie)
      If (the Romanized Arabic character(ia) = the English character(ie))
            SIM1 = SIM1 + 1
      End
      Else ie = ie + 1 & Read the English character(ie)
            If (the Romanized Arabic character(ia) = the English character(ie))
                  SIM1 = SIM1 + 1
            End
            Else If(English character(ie - 2)= English character(ie - 1)))
                  ie = ie+1 & Read the English character(ie)
                  If (the Romanized Arabic character(ia) = the English character(ie))
                        SIM1 = SIM1 + 1
                  End
            End
      ia = ia + 1 & ie = ie + 1
      if (ia < Length (Romanized Arabic word))
            GOTO R
      End
SIM1 = SIM1/(max_Length(Romanized Arabic word, English word))
```

17.2.3 Similarity measure 2 (SIM2)

Using SIM2, the system restricts the extracted transliteration pairs only to the pairs that have all romanized Arabic characters matched with some or all English proper noun characters to increase the precision. We achieve that by modifying the algorithm mentioned in Section 17.2 to set the similarity score to zero if any romanized Arabic character does not match with any English character. Therefore, for using any threshold value (th) > 0, the transliteration pairs that do not have all romanized Arabic characters matched with some or all English proper noun characters are excluded.

17.2.4 Similarity measure 3 (SIM3)

Arabic vowels and diacritics exist extensively in the Arabic text but do not appear on words in almost all Arabic documents. Therefore, considering these vowels and diacritics when the system specifies the transliteration pair matching score value will decrease it. Modifying the algorithm mentioned in Section 17.2 so that if the algorithm found a certain short vowel between two successive consonants (‘a’ = “َ”, ‘o’ = “ُ”, ‘i’ = “ِ”, ‘e’ = “ِ”) or a repeated consonant (“ّ”), the algorithm does not consider them at all. Hence, the only modification will be in the last step of the algorithm to specify the new similarity value SIM3 as follows:

SIM3 = SIM3/(Max_Length(romanized Arabic Word, English Word(after excluding short vowel between two successive consonants and repeated consonant))).

17.3 Experimental results

We have applied our transliteration techniques on the “Arabic English Parallel News Text Part 1,” Linguistic Data Consortium (LDC) catalog number LDC2004T18 and ISBN 1-58563-310-0. This corpus contains Arabic news stories and their English translations LDC collected via Ummah Press Service from January 2001 to September 2004. It totals 8,439 story pairs, 68,685 sentence pairs, 2M Arabic words, and 2.5M English words (http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2004 T18). This corpus contains 8827 English proper nouns specified by using the CLAWS4 POS tagger.

17.3.1 Experimental results using pattern matching approach

We treat the transliteration compilation problem as a pattern matching problem in [7]. We applied the approach on the English–Arabic corpus. We achieved precision and recall of 68.3 and 67.4%, respectively, for the best matched pairs. We also achieved precision and recall of 71.6 and 69.7%, respectively, for the top three Arabic transliterations for an English proper noun, respectively. Then we did little modification in the first step of the approach to increase precision. In the first step of the algorithm, we did not tag the English half of the parallel text only we also tagged the Arabic half in order to restrict the matching process to as few words as possible to increase precision. We achieved precision and recall of 71.4 and 66.5%, respectively, for the best matched pairs. We also achieved precision and recall of 73.8 and 68.2%, respectively, for the top three Arabic transliterations for an English proper noun, respectively. We found that many mistaken transliterations resulted from insufficient data.

17.3.2 Experimental results using algorithms 1 and 2 combination of Fattah et al. [6]

We have applied algorithms 1 and 2 combination of Fattah et al. on the Arabic English Parallel News Text Part 1 corpus after stemming and preprocessing steps that were mentioned in Fattah et al. We achieved precision of 89.3% and recall of 74.6%. The precision is better than that of Fattah et al. However, the recall is deteriorated. The reason of low recall is that many proper nouns are compound nouns such as the Arabic proper name "عبد القادر" and the English transliteration of it "Abdel Qader." The first algorithm of Fattah et al. fails to capture dependencies between groups of words and the second algorithm of Fattah et al. can extract the word and phrase translation of one word only. Hence, algorithms 1 and 2 combination failed to extract the compound nouns such as "عبد القادر" and "Abdel Qader" which deteriorates the recall.

17.3.3 Experimental results using algorithms 1 and 2 combination of Fattah et al. [6]

We have applied the proposed approach using Dice's similarity coefficient on the English–Arabic parallel corpora to extract all possible transliteration pairs.

Apply the previous pseudo-code on the English–Arabic corpora and use Dice's smilarity coefficient to measure the similarity between the English proper noun and the romanized Arabic proper noun. We extract all transliteration pairs of similarity scores that exceed the threshold “th” value. Table 17.1 shows the precision, recall, and the harmonic mean of precision and recall (F-Measure) for the transliteration pairs extracted as a function of the threshold “th.”

As shown in Table 17.1, the precision is high especially for “th ≥ 0.7”; however, the recall is too small in all values of “th” except “th” = 0.0. As “th” decreases, the precision decreases and the recall increases as well. For “th” = 0.0, the system creates all possible transliteration pair combinations included in a sentence pair. Hence, at “th” = 0.0, the precision is minimum while the recall is maximum. At “th” = 0.0, the recall is not 100% since there is some error caused by the English tagger and the Arabic morphological analyzer tools. The precision and recall at “th” = 0.0 are fixed for any similarity measure used. From Table 17.1 we can see that the recall is low in general since Dice’s approach is not the best way to measure similarity between an English proper noun and the romanized Arabic proper noun. That is because the Arabic diacritics are not typed in most of the Arabic documents. Take the following example to illustrate this point.

The Arabic proper noun “محمد = Mohammad” is written in a very few documents (such as Muslim holy book) as “مُحَمَّد” whereas it is written as “محمد” without any short vowel or diacritic in almost all other documents. The diacritics appearing on the Arabic proper noun “محمد” are ‘a’ = ‘َ’, ‘o’ = ‘ُ’, and a repeated consonant ‘m’ that is considered as a diacritic ‘ّ’. The correct transliteration of “مُحَمَّد” is “Mohammad,” where the romanization of “محمد” is “mhmd.” If we use Dice’s approach to measure the similarity between “Mohammad” and “mhmd,” the score will be 0 (SIM(“mohammad,” “mhmd”) = 2*0/(3+7) = 0). Hence many correct transliteration pairs have low Dice’s similarity coefficient value that forces the system to discard them.

Table 17.1. The results using Dice’s similarity coefficient

th	1.0	0.9	0.8	0.7
Precision	100%	100%	95.9%	86.7%
Recall	2.3%	2.3%	6.2%	15.3%
F-Measure	4.5%	4.5%	11.6%	26.0%
th	0.6	0.5	0.3	0.0
Precision	72.1%	61.3%	42.8%	24.2%
Recall	22.6%	28.7%	36.5%	98.1%
F-Measure	34.4%	39.1%	39.4%	38.8%

17.3.4 Experimental results using SIM1

As illustrated in the previous section, Dice's approach to measure similarity between English–Arabic transliteration pairs does not work well. It leads us to use another approach of similarity called "SIM1" as illustrated in Section 17.2.

Apply the algorithm of Section 17.2 on the transliteration pair "mhmd = محمد, mohammed," SIM1 = 4/8 = 0.5 instead of 0 in the case of using Dice's approach. Hence, if the threshold "th" = 0.5, the transliteration pair "mhmd = محمد, mohammed" will be included in the final file. Table 17.2 shows the results when we apply the algorithm in Section 17.2 to specify SIM1 as a similarity score for the transliteration pair under consideration.

It is clear from Table 17.2 that the recall has been improved compared with Table 17.1. However, the precision is slightly decreased. For th = 1, all the romanized Arabic characters are matched such as "alahram, الأهرام" (the romanization form of "الأهرام" is "alahram"), "mark, مارك", and "taba, طابا."

For th = 0.9, almost one character in a long word is not mapped properly, such as "kazakhistan, كازاخستان = kazakhstan," only the short vowel 'i' does not appear in the Arabic word. The same word "kazakhistan" appears in th = 1 as "kazakhstan, كازاخستان."

Table 17.2. The results using SIM1 similarity coefficient

th	1.0	0.9	0.8	0.7
Precision	100%	100%	92.4%	75.7%
Recall	2.3%	6.5%	26.7%	45.2%
F-Measure	4.5%	12.2%	41.4%	56.6%
th	0.6	0.5	0.3	0.0
Precision	61.8%	36.3%	22.1%	24.2%
Recall	57.1%	62.4%	78.7%	98.1%
F-Measure	59.4%	45.9%	34.5%	38.8%

For th = 0.8, most of the errors occurred with short words that have matching score equal to or more than 0.8. For instance in the transliteration pair, "aladl, الامل = alaml," four characters are matched and this transliteration is not correct. It is better to match all consonant and long vowel characters of the converted word to avoid this kind of error. Examples: "alalam, الإسلام = alaslam." Another problem is "salam, سالم = salm," when, all the converted word characters are matched, the transliterated pair is not correct. Some others have more than one transliteration and all are correct such as "tahir, طاهر = tahr, taher." This occurs since the Arabic language has only three short ('a', 'e', 'o') and three long ('a', 'y', 'w') vowels.

Hence the language does not differentiate between the English vowels ‘i’ and ‘e.’ Another example occurred because of different pronunciation of the people such as “alsobah, الصباح = alsbah, alsabah.” For th = 0.5, the correct pairs are few such as “Mohammad, محمد.”

17.3.5 Experimental results using SIM2

As we notice in the previous section, in the transliteration pair “aladl, الامل,” when the Arabic word “الامل” is converted to English alphabet, it will be “alaml.” If we match “aladl” with “alaml,” only ‘d’ and ‘m’ do not match. So the similarity score is SIM1 = 0.8. And the pair is not correct. Hence, it is required that the system restricts the extracted transliteration pairs only to the pairs that have all romanized Arabic characters matched with some or all English proper noun characters to increase the precision. We achieve that by modifying the algorithm mentioned in Section 17.2 to set the similarity score to zero if any romanized Arabic character does not match with any English character. Hence we use a new similarity measure called SIM2. SIM2 = 0 if any romanized Arabic character does not match with any English character. Using SIM2 as a similarity measure, we achieved the results in Table 17.3.

Table 17.3. The results using SIM2 similarity coefficient

th	1.0	0.9	0.8	0.7
Precision	100%	100%	98.9%	94.5%
Recall	2.3%	6.5%	24.6%	42.3%
F-Measure	4.5%	12.2%	39.4%	58.4%
th	0.6	0.5	0.3	0.0
Precision	88.6%	56.1%	34.2%	24.2%
Recall	53.4%	60.3%	66.4%	98.1%
F-Measure	66.6%	58.1%	45.1%	38.8%

The drawback of this restriction is that the recall is slightly decreased. Although the precision is increased in general, there are some errors. For instance, “hamdi, حامد = hamd.” However, when romanized Arabic characters are matched with some English characters, the transliteration pair is not correct.

17.3.6 Experimental results using SIM3

In the example, “Mohammad, محمد = mhmd,” the score SIM2 = 0.5. If th is higher than 0.5, this transliteration pair will be discarded and consequently it decreases the recall. It is well known that short vowels and diacritics do not appear on the Arabic word. For instance, "مُحَمَّد" is written as "محمد" without any short vowels or diacritic. Hence, we consider SIM3 as in Section 17.4.

In the previous example "Mohammad, محمد = mhmd." If we applied the algorithm in Section 17.2 after modification, the new similarity score "SIM3" will be 4/4=1 instead of 0.5. Applying this approach, and do not restrict the extracted transliteration pairs only to the pairs that have all romanized Arabic characters matched with the English proper noun characters, we achieved the results in Table 17.4.

Table 17.4. The results using SIM3 similarity coefficient

th	1.0	0.9	0.8	0.7
Precision	99.3%	99.1%	96.3%	87.5%
Recall	25.6%	26.0%	56.4%	68.2%
F-Measure	40.7%	41.2%	71.1%	76.7%
th	0.6	0.5	0.3	0.0
Precision	81.2%	51.1%	31.1%	24.2%
Recall	76.9%	77.3%	83.4%	98.1%
F-Measure	79.0%	61.5%	45.3%	38.8%

17.4 Conclusions

In this study, we presented a new system to create English–Arabic transliteration pairs from parallel corpora based on different similarity measure approaches. The strength of our new system is that it works well for low-frequency transliteration pairs. The system could extract some correct transliteration pairs of frequency equal to 1. We found that the similarity measure must be specified based on the characteristics of the two language pairs under consideration. We have evaluated the presented new approaches using the English–Arabic parallel corpora. Most of our results outperform previously published results in terms of precision, recall, and F- Measure. We believe that the presented approach will improve the precision and recall in cross-language information retrieval system.

In the future work, we will use the resulted transliteration pairs in cross-language information retrieval and machine translation systems.

Acknowledgment. This research has been partially supported by the Japan Society for the Promotion of Science (JSPS), Grant No. 07077, and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (B), 19300029.

References

1. Al-Onaizan Y, Knight K (2002) Translating named entities using monolingual and bilingual resources. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, pp 400–408
2. Stalls B, Knight K (1998) Translating names and technical terms in Arabic text. In: Proceedings of the COLING/ACL Workshop on Computational Approaches to Semitic Languages
3. Chen HH, Huang SJ, Ding YW, Tsai SC (1998) Proper name translation in cross-language information retrieval. In: Proceedings of 17th COLING and 36th ACL, pp 232–236
4. Knight K, Graehl J (1998) Machine transliteration. Computational Linguistics 24(4):599– 612
5. Kang BJ, Choi KS (2001) Two approaches for the resolution of word mismatch problem caused by English words and foreign words in Korean information retrieval. International Journal of Computer Processing of Oriental Languages 14(2):109–131
6. Fattah M, Ren F, Kuroiwa S (2006a) Stemming to improve translation lexicon creation form bitexts. Information Processing & Management 42(4):1003–1016
7. Fung P (1995) A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. CoRR cmp-lg/9505016
8. Fung P, Yee L (1998) An IR approach for translating new words from nonparallel, comparable texts. COLING-ACL, pp 414-420
9. Fung P (1998) A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. AMTA (1998), pp 1-17
10. McEnery AM, Oakes MP (1996) Sentence and word alignment in the crater project. In: Thomas J, Short M (eds) Using Corpora for Language Research, Longman, London, pp 211–231

Chapter 18

Heart rate variation adaptive filtering based on a special model

S. Seyedtabaii

Electrical Engineering Department, Shahed University, Tehran, Iran,
(E-mail: stabaii@shahed.ac.ir)

Abstract. The low- and high-frequency components of an RR heart rate signal must be adequately separated to provide accurate heart rate variability indices of sympathetic and parasympathetic activity. Adaptive filters can separate the low-frequency sympathetic and high-frequency parasympathetic components from an ECG RR interval signal, enabling the attainment of more accurate heart rate variability measures. For the raised case, this chapter suggests an efficient (short size) case-based model and illustrates its performance in adaptive filtering of heart rate signal. This method renders analogous results to what a higher order conventional FIR model adaptive filter may yield. The advantage of this model lays in its ability to accommodate the phase difference between breathing signal and the HF component of HRV using a low-order tunable filter. Simulation results supporting the proposed scheme are presented.

Keywords. Adaptive filter, FIR model, First-order equalizer, HRV filtering

18.1 Introduction

Heart rate variability (HRV) is a measure of alterations in heart rate derived by measuring the variation of RR intervals. HRV parameters have been shown to aid assessment of cardiovascular disease [1]. Heart rate is

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_18,

influenced by both sympathetic and parasympathetic (vagal) activity. The influence and balance of both branches of the autonomic nervous system (ANS) have been termed sympathovagal balance and is reflected in the RR interval changes. A low-frequency (LF) component of HRV has been proposed as reflecting both sympathetic and parasympathetic effects on the heart and generally occurs in a band between 0.04 and 0.15 Hz. The influence of vagal efferent modulation of the sinoatrial node can be seen in the high-frequency band (HF), loosely defined between 0.15 and 0.4 Hz and known as respiratory sinus arrhythmia (RSA) because it occurs at the respiratory frequency. The magnitude of this high-frequency band has been demonstrated to be associated with the extent of cardiac parasympathetic activity in pharmacological autonomic blockade studies [2], respiratory sinus arrhythmia, cardiac vagal tone, and respiration within and between-individual relations.

The ratio of power in the LF and HF components (LF/HF) has been used to provide an estimate of cardiac sympathovagal balance, although this measure remains indisputable [3]. Nevertheless, several studies have indicated that when considered jointly, HF and LF HRV may provide useful information about both sympathetic and parasympathetic influences upon the cardiac cycle [4].

Spectral HRV is a measure of power in various frequency bands. To determine the RSA amplitude over a period of time, frequency domain, time domain, and phase domain approaches have been analyzed [5]. Presented in [6] is an adaptive filter that separates the LF and HF components and therefore yields distinct spectral analysis measures for each band. The suggested order for the used FIR filter is 20. In this chapter an adaptive filter with a new model structure, with just a few parameters, is introduced which behaves similar to the higher order FIR model adaptive filter in facing RSA filtering.

In Section 18.2 the high-frequency component of HRV signal is briefly analyzed. Section 18.3 embodies a short description of adaptive filter and in Section 18.4 the basis of the proposed model is presented. Section 18.5 contains simulation results and finally conclusion comes in Section 18.6.

18.2 HF HRV signal analysis

18.2.1 RSA frequency

Simultaneous oscillation of heart rate (HR) and blood pressure (BP) at the breathing frequency was first observed by Hales in 1733. The respiration-related fluctuation of HR has been named “respiratory sinus arrhythmia” (RSA), and it manifests as increasing HR upon inspiration and decreasing HR upon expiration [7]. On the other hand, parallel oscillation of RR intervals with nerve activity in the absence of lung movements have also been reported [8] that indicates the association between the central respiratory drive and respiratory-related cardiovascular oscillation, thus, it would be wise to regard HRV oscillation even in the absence of lung movement.

Related to the importance of respiration, the logical conclusion is that once the actual breathing rate is known, detection of the HF power should be centered around the fundamental respiration frequency and not a default fixed frequency which is the case with traditional HRV analysis. This also implies that breathing pattern (V_t) is a good signature for removal of HF component from HRV.

18.2.2 RSA phase shift

The results of a synchronized respiration-heart beat experiment shows that there is a variable phase difference (delay) between HF HRV signal and the respiratory signal (V_t) that changes when subject shifts from sitting position to supine position [9]. Vaschillo et al. [10] also measure the frequency response of HF HRV versus the stimulus (breathing) and show that phase varies monotonically with frequency and its value is approximately 0 at 6 min^{-1}.

18.3 Conventional adaptive filtering

In order to separate the LF and HF components of an RR interval signal, prior to spectral analysis, the RR interval (R_R) and the tidal volume (V_t) signals are applied to an adaptive filter with FIR model shown in Fig. 18.1. In this, $H(z)$ is an FIR tunable filter as follows:

$$y(k) = \sum_{i=0}^{N-1} w_i V_t(k-i) \tag{18.1}$$

where $W=[w_0,\ldots,w_{N-1}]$ is its parameter vector, $V_t(k)$ is input, and $y(k)$ is its output. In the HRV adaptive filtering, the parameters of H(z) are to be tuned by an algorithm such as LMS in such a way that the output $y(k)$ resembles to RSA, the trace of $V_t(k)$ in R_R. This happens when the mean square error between $y(k)$ and RSA (laid in R_R) is minimized by computing optimally the filter coefficients.

In LMS, the weights are updated on a sample-by-sample basis as follows:

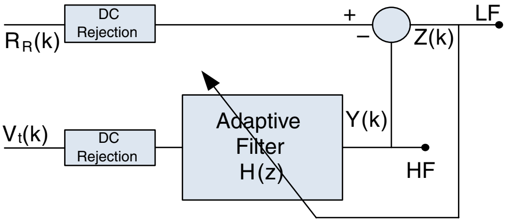


Fig. 18.1. Adaptive filter with FIR model

$$W_i(k+1) = W_i(k) + 2\mu V_t(k-i)\left[R_R(k) - y(k)\right] \qquad i = 0,\ldots,N-1 \quad (18.2)$$

where N is the filter order. This is a practical approach to obtaining estimates of the filter weights (W) in real time without having to perform extensive computations. The algorithm does not require prior statistical knowledge of the signal and instead uses instantaneous estimates. Therefore, the weights obtained by the LMS algorithm are estimates that gradually improve over time as the filter weights are adjusted as the filter learns the characteristics of the signal and eventually converge.

In the implementation, the set of weights is first initialized to zero. Then, for each subsequent sampling instants k, the filter output is computed using the FIR filter expressed by Eq. (18.1), where the output $y(k)$, predicting the respiratory or RSA component in the RR interval signal, is the filtered tidal volume (V_t). Having the predicted RSA, it is now possible to linearly subtract it from RR interval signal, R_R. The error estimate is the algorithm output and is computed by

$$z(k) = R_R(k) - y(k)$$

where $z(k)$ is the LF component and $y(k)$ is the predicted HF or RSA component. The filter weights W are updated based on this error expressed by Eq. (18.2), where μ controls the rate of convergence and the stability of the algorithm.

18.4 The proposed model

Having a close look at the oscillatory nature of the two signals, $V_t(k)$ and $R_R(k)$, it guides us to a better model structure for adaptive filter. If it is assumed that most of the $V_t(k)$ power is in its fundamental frequency, as it is the case, it can be modeled by a cosine function,

$$V_t(k) = A\cos(\omega k)$$

Then, its trace in the RR interval, $R_{SA}(k)$, will also be a type of shifted cosine with certain amplitude,

$$R_{SA}(k) = B\cos(\omega k - \varphi)$$

For exact cancellation of $R_{SA}(k)$ from $R_R(k)$, $y(k)$ has to be,

$$y(k) = B\cos(\omega k - \varphi)$$

In this chapter, for generating the required $y(k)$ out of $V_t(k)$, the method of summation of $V_t(k)$ with its arbitrary shifted version is suggested as follows:

$$\alpha_1[A\cos(\omega k)] + \alpha_2[A\cos(\omega k - \delta)] = B\cos(\omega k - \varphi)$$

This can also be written in vector form

$$\alpha_1[Ae^{j0}] + \alpha_2[Ae^{-j\delta}] = Be^{-j\varphi}$$

The solution to this equation is,

$$\alpha_1 = \frac{B\sin\varphi}{A\sin\delta} \qquad \alpha_2 = \frac{B\sin(\delta - \varphi)}{A\sin\delta}$$

The solution does not set a specific value for δ, except that it must be nonzero. One conclusion may be that a first-order FIR model adaptive filter (which has two parameters) can also be able to filter the signal. But it never happens. Even FIR filter with order 8 has difficulty in successful removal of RSA. The reason is that the choice of δ noticeably determines the rate of convergence of the underlying LMS algorithm, when the LMS algorithm for online optimal tuning of the filter parameters is used. This is the advantage of the proposed method, which this chapter tries to exhibit.

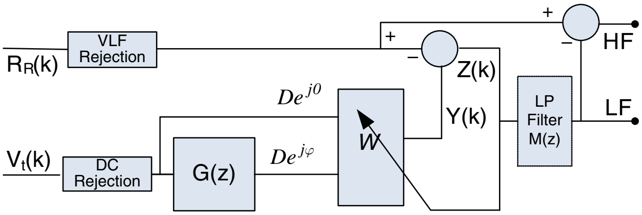


Fig. 18.2. Adaptive filter with a special model

This type of model can also be expressed in geometrical concept. A cosine function can be represented by a vector, where its angle equals the cosine phase shift. Since in a plane, any vector can be formed by summation of two out-of-phase variable length vectors (with probable minus sign), any shifted phase cosine can also be generated by summation of two different phase-shifted cosine.

The way that this idea is accommodated in adaptive filtering is shown in Fig. 18.2. In this structure, input V_t is injected into the algorithm through two branches. From the first branch, it directly enters and forms vector 1, $V_t=De^{j0}$, that is, in phase with V_t. From the second branch, vector 2, $V_{td}=De^{j\varphi}$, is formed, that is, a phase-shifted version of V_t. To do so, V_t is passed through a known allpass filter, $G(z)$

$$G(z)=\frac{z^{-1}-\beta}{1-\beta z^{-1}}$$

This filter has unity amplitude and inserts necessary phase shift in the input signal, so that $V_t = De^{j0}$ is transformed to $V_{td} = De^{j\varphi}$. Then the two branches enter the tunable parameter block and are summed together.

$$y = w_0 V_t + w_1 V_{td}$$

This output provides the desired shifted version of V_t needed for subtraction from RR interval to remove RSA signature. This is achieved, once weights have got properly adjusted.

The choice of β alters the shape of cost function hyperspace an manages the rate of convergence of the LMS algorithm. Experiments support this proposition. Note that by setting β=0, the proposed scheme turns to the conventional adaptive filter. Therefore, β has to be set appropriately. Fortunately, in this case, sensitivity of the algorithm to the value of β is low. For respiratory signal frequency between 0.15 and 0.4 Hz, real-world span of the signal, and under various phase shifts, a value between 0.6 and 0.9 for β can fulfill the job. Search for an optimal value for β can easily be embedded in the LMS algorithm, but for this case is not really needed.

18.5 Simulations

18.5.1 Experiment 1

In this experiment, harmonics of the signals are ignored or assumed to have already been filtered.

The tidal volume data may be collected from the LifeShirt. The LifeShirt contains two inductive plethysmography (IP) sensors encircling the ribcage and abdomen used to measure tidal volume. Tidal volume simulation signal [6] can be

$$Vt(t) = D\cos(2\pi f_h t) + E \tag{18.3}$$

where tidal volume is expressed by *D,* and *E* is a DC offset whose values depend on ribcage and abdominal cross-sectional area, which varies with changes in posture and mass. The LifeShirt also contains a single lead ECG sampled at 200 Hz, which is linearly interpolated to 1 kHz, and heart rate is determined based on R wave locations. The RR interval signal for simulation may be represented by [6]:

$$s(t) = A\sin(2\pi f_h t + \alpha_h) + B\sin(2\pi f_l t + \alpha_l) + C \tag{18.4}$$

where *A* is the peak-to-peak RSA amplitude per breath, expressed in msec, *B* is the LF/HF ratio, expressed as a fraction of *A,* and *C* is the mean RR interval. The parasympathetic (HF) component and sympathetic (LF) component have frequencies f_h, f_l, and phases α_h, α_l, respectively. Based on the discussion in Section 18.2, RSA component of R_R has the same frequency as V_t with a (variable) phase difference. Both signals are assumed to be sampled in 5 ms intervals.

18.5.1.1 FIR Model adaptive filter

The simulated RR interval signal is illustrated in Fig. 18.3 for a parameter set with high and low frequencies of 0.18 and 0.15 Hz, respectively, where power leakes from the LF band into the adjacent HF band.

No phase variation is applied to this signal. An RSA amplitude of $A = 200$ msec is used with B=100 to give a 50% LF/HF ratio. These signals are applied to the adaptive filter. Figure 18.4 shows the predicted HF component within the RR signal, predicted-based on the reference signal V_t. It is evident from the trace that it takes approximately 50 s for the filter to tune and adapt to simulation characteristics. The weighting parameters are as follows:

w = [–0.0542, −0.0093, 0.0361, 0.0797, 0.1192, 0.1526, 0.1783, 0.1949, 0.2015, 0.1979]

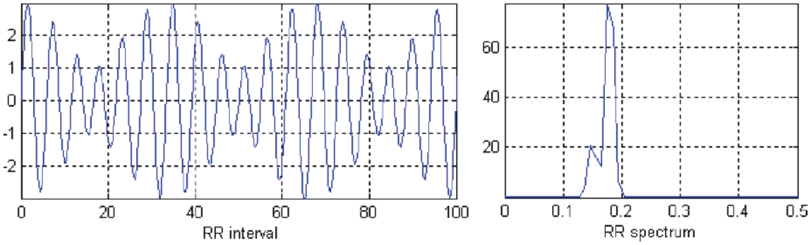


Fig. 18.3. Simulated RR interval signal and its spectrum

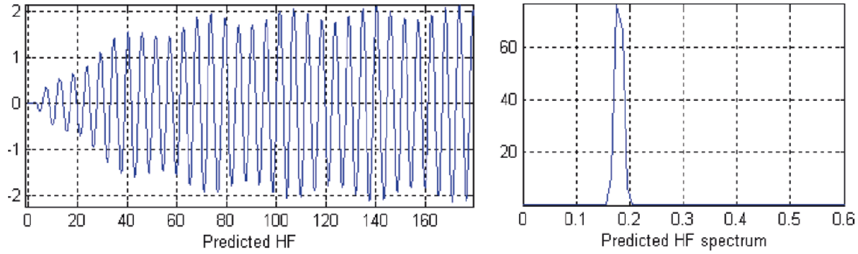


Fig. 18.4. Predicted HF and its spectrum using FIR model with N=10

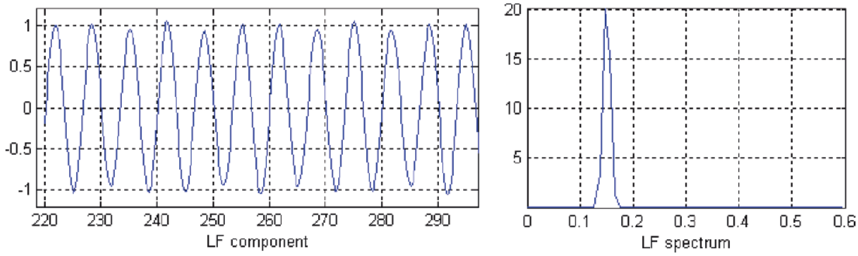


Fig. 18.5. Predicted LF and its spectrum using FIR model with N=10

The LF signal is derived by linearly subtracting the HF signal from the original raw signal. Figure 18.5 shows the separated LF signal and its spectrum. It is obvious that the HRV has been accurately decomposed. These results are obtained with filter order of N=10, after removal of the RR interval and V_t means. Increasing the filter order does not improve the results as is the case with LMS. Decreasing it below $N=6$ leads to algorithm complete failure. The step size parameter, $\mu=6*10^{-4}$ produced the best result. This value appears to be very small, as the input signals have not been normalized. This is one drawback using the ordinary LMS adaptive filter, since when heart rate and respiratory amplitude vary the updating parameter requires retuning. This is resolved with the normalized least mean squares (NLMS), which is common in most software packages. This approach has been shown to increase accuracy when applied to current HRV spectral analysis techniques. However, when applying linear subtraction, although the predicted signal may be nearly perfect, any slight phase variation creates large artifact in the resultant signal [6].

This result is the reproduction of the investigation reported in [6], done with FIR model of order N=20. Apparently, the reason for using N=20, as we noticed in our simulations, is due to the DC offset of V_t signal that probably had not got removed.

18.5.1.2 The proposed model adaptive filter

This time the simulated *RR* interval and $V_t(k)$ signal are applied to the proposed model with β=0.8. In this simulation the LP filter, M(z), is not needed and can be ignored, since in this experiment the signals are assumed to be free of harmonics. Figures 18.6 and 18.7 show the results. The optimum filter parameters are W=[*0*.1353, 1.0316]. In the simulation, the RR interval and V_t signals' average have been removed. The acceptable result with μ=*5**10^{-3} is obtained.

As the graphs show, the proposed model having just two parameters works analogous to the FIR filter with *N=20* as reported in [6] or FIR filter with *N*=10 that we used for the reproduction of the results of [6].

What is important with this model, as our experiments indicate and the model structure suggests, is that this model is able to strongly tackle the phase shift variation in α_h, what the conventional adaptive filter fails to accomplish easily. This is, of course, true for the encountered case and the capacity of this model to manage other situations has to be investigated.

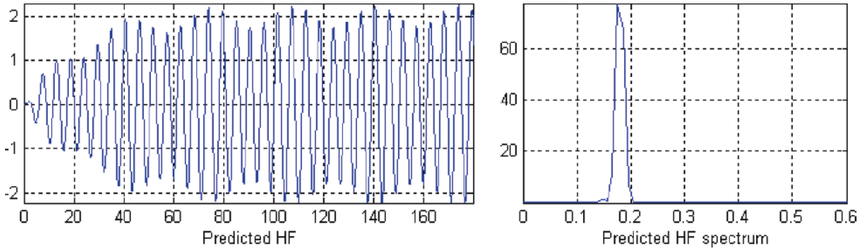


Fig. 18.6. Predicted HF and its spectrum using the proposed model with *N*=2

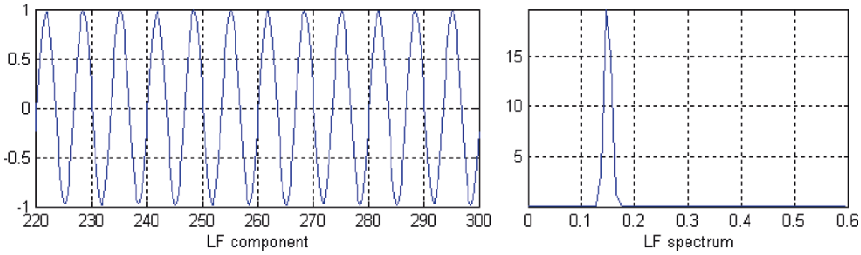


Fig. 18.7. Predicted LF and its spectrum using the proposed model with *N*=2

18.5.2 Experiment 2

In this experiment, the first harmonic of the signals is also included. Figure 18.8 depicts the $V_t(k)$ and its spectrum showing breathing fundamental frequency and its first harmonic. In this simulation, for test purposes, the frequency of the LF is 0.17 Hz and RSA frequency is given as 0.15 Hz, which is in band with the LF frequency. Since RSA frequency is variable and is in band with the LF frequency, the traditional method of filtering cannot be pursued, instead adaptive filtering is a valuable choice.

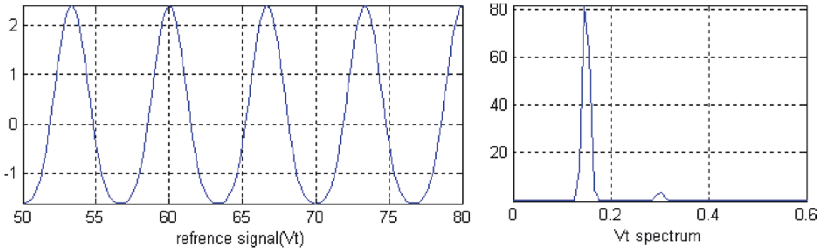


Fig. 18.8. $V_t(k)$ and its spectrum in experiment 2, containing a 0.15 Hz and its 0.3 Hz first harmonic

18.5.2.1 FIR model adaptive filter

Figure 18.9 is the output of LMS adaptive filtering with N=8 and Fig.18.10 is its output by filter order of 10. The result with N=8 shows failure of LMS in removing $R_{SA}(k)$ from $R_R(k)$ appropriately.

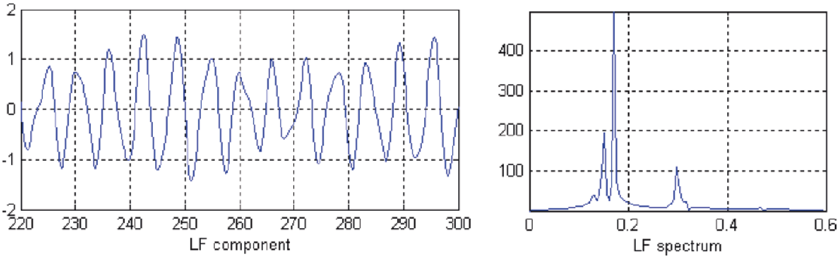


Fig. 18.9. Predicted LF and its spectrum in experiment 2: LMS with N=8

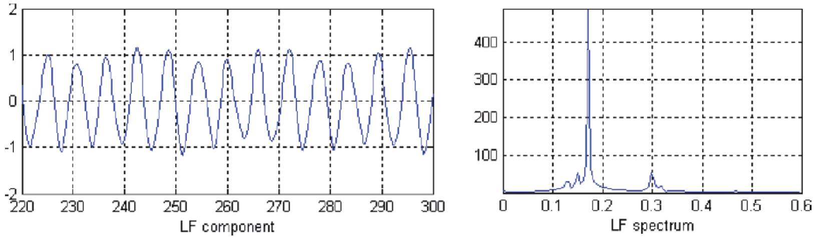


Fig. 18.10. Predicted LF and its spectrum in experiment 2: LMS with N=10

18.5.2.2 The proposed model adaptive filter

This time the simulated RR interval and $V_t(k)$ signal are applied to the proposed model with β=0.8. The LP filter, M(z), used for harmonic rejection is a second-order elliptic filter. The passband edge of the filter is the LF HRV upper limit and stopband edge is the frequency of the first harmonic of the HF HRV lower limit, 0.3 Hz. Using a fixed filter for harmonic rejection is a good choice, since harmonics of $V_t(k)$ and $R_{SA}(k)$ are no longer in band with the LF frequency. With this provision, the load of online tuning still remains very low, just adjusting two parameters. Figure 18.11 shows the result. The optimum filter parameters are w=[−0.3282*,* 1.0364]. As the graphs show, the performance of the proposed model has nothing less than what the FIR model with filter order of 10 yields, besides its ability to track the variability of the phase shift of RSA is an important asset.

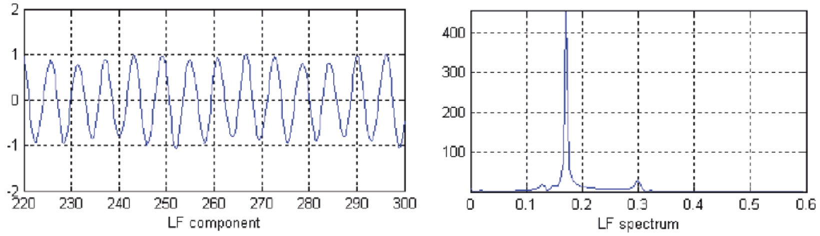


Fig. 18.11. Predicted LF and its spectrum in experiment 2: the proposed method with N=2

18.6 Conclusion

In this chapter a new model structure for adaptive filtering, based on the nature of the involved signals, is introduced that gives similar performance as what a higher order FIR model adaptive filter in removing RSA component of HRV may yield. The trace of V_t signal in RR interval (RSA) is an unknown shifted phase of V_t with, generally, unknown amplitude. On this basis, the proposed model is designed with the capability of tracking the phase-shifted breathing pattern in the RR interval. This is accomplished by summing together V_t signal with its arbitray shifted version through a set of optimally adjusted weights. The algorithm is not too much sensitive to the variation of amounts of shift; however, the appropriate value of shift improves the convergence rate of the algorithm as the simulation results indicate.

Acknowledgment. This work was partially supported by ETRC, Shahed university, Tehran, Iran.

References

1. Crawford MH, Bernstein S, Deedwania P (1999) ACC/AHA guidelines for ambulatory electrocardiography. Circulation 100:886–893
2. Grossman P, Kollai M (1993) Individual differences in respiratory sinus arrhythmia, intra individual variations and tonic parasympathetic control of the heart. Psychophysiology 30:486–495
3. Eckberg DL (1997) Sympathovagal balance: a critical appraisal. Circulation 96:3224–3232
4. Grossman P, Van Beek JA (1990) A comparison of three quantification methods for estimation of respiratory sinus arrhythmia. Psychophysiology 27:702–714
5. Pagani M, Montano N, Porta A (1997) Relationship between spectral components of cardiovascular variabilities and direct measures of muscle sympathetic nerve activity in humans. Circulation 95:1441–1448
6. Keenan DB, Grossman P (2005) Adaptive filtering of heart rate signals for an improved measure of cardiac autonomic control. Int J Signal Process 2:52–58
7. Sasano N, Vesely AE, Hayano J, Sasano H, Somogyi R, Preiss D et al. (2002) Direct effect of PaCO2 on respiratory sinus arrhythmia in conscious humans. Am J Physiol Heart Circ Physiol 282:973–976
8. Shykoff BE, Naqvi SS, Menon AS, Slutsky AS (1991) Respiratory sinus arrhythmia in dogs: Effects of phasic efferents and chemostimulation. J Clin Invest 87(5):1621–1627
9. Robert P, Daniel K (1997) Heart rate change as a function of age, tidal volume and body position when breathing using voluntary cardiorespiratory synchronization. Physiol Meas 18:183–189
10. Vaschillo E, Vaschillo B, Lehrer P (2004) Heartbeat synchronizes with respiratory rhythm only under specific circumstances. Chest 126:1385–1387

Chapter 19

Performance evaluation of table-driven and buffer-adaptive WLANs

Imam Al-Wazedi[1], A. K. Elhakeem[2]

Department of ECE, Concordia University, H3G 2W1, Canada
[1]i_alwaze@encs.concordia.ca, [2]ahmed@ece.concordia.ca

Abstract. The random access control (MAC) technique of standard WLANs is called the distributed coordination function (DCF) [3]. DCF is a carrier sense multiple access based on collision avoidance (CSMA/CA) scheme with binary slotted exponential backoff. This exponential backoff makes the system more complex, and fairness [3−13] among the stations is a major concern. This chapter shows one possible evolution of WLANs where exponential backoff is not employed. In the new techniques herein users transmit randomly but adapt themselves to traffic conditions, thus improving throughput and delay while guarantying fairness. Users in the first technique are controlled by a table which is derived from traffic measurements (table driven). In the second, users' transmission activities are function of their buffer contents.

Keywords. MAC, Table-driven WLANs, Buffer-adaptive WLANs

19.1 Introduction

Wireless local area networks (WLANs) have been widely deployed for the past decade. Their performance has been the subject of intensive research. In [1] improvement of throughput and fairness is shown by optimizing the backoff. Xuejun et al. [1] use a measure called the average idle interval which does not consider the number of collisions. In [2], the authors proposed a MAC layer-based WLAN technique in which they gave higher

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_19,

priority to access the channels so as to improve the throughput and the channel utilization. Xuejun et al. and Liang et al. [1, 2] discuss the fairness problem of the exponential backoff. Bharghvan [3] proposes a technique based on collision avoidance and fairness to improve the channel utilization. Few WLAN standards have been adopted, e.g., IEEE 802.11 [4] which uses collision avoidance scheme with binary slotted backoff. Bianchi [4] expresses how the throughput deteriorates with increasing the number of nodes. Tay and Chua [5] use an analytic model to study the channel capacity – i.e., maximum throughput – when using the basic access (two-way handshaking) method. Kim and Lee [6] consider three kinds of CSMA/CA protocols, which include basic, stop-and-wait, and four-way handshake CSMA/CA, and introduce a theoretical analysis for them. Cali [7] pointed out that depending on the network configuration, DCF may deliver a much lower throughput compared to the theoretical limit. In [8] a contention-based MAC protocol named fast collision resolution is presented. Wu et al. [10] propose a model named DCF+ which shows the fairness performance. Chnaya and Gupta [11] present the performance evaluation of the decentralized nature of communication between nodes in IEEE 802.11, in the presence of "hidden" nodes. Estiaghi et al. [13] show the performance evaluation of multihop ad hoc WLANs. Thus extensive research has been conducted on WLANs [1–13]. Fairness index was only discussed in [1], [2], [7], and [10].

This chapter tries to investigate simultaneously the four performance indexes, i.e., throughput, delay, delay variance, and fairness, which are not considered in previous studies [1–13]. In the table-driven technique we consider both idle periods and number of collisions (Table-driven technique) which show the actual load on the network. In the second, we employ buffer-adaptive technique which guaranties fairness and provides smaller delay variance.

19.2 The IEEE 802.11 MAC protocol

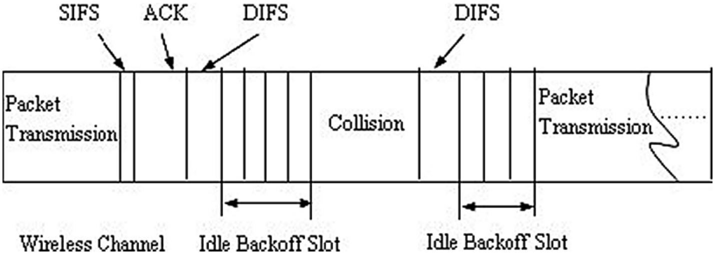


Fig. 19.1. IEEE 802.11 MAC mechanism

Figure 19.1 shows one of many transmission scenarios possible with the IEEE 802.11 DCF mode. In this mode a node with a packet to transmit initializes a backoff timer with a random value selected uniformly from the range [0, CW-1], where CW is the contention window in terms of time slots. After a node senses that the channel is idle for an interval called DIFS (DCF interframe space), it begins to decrease the backoff timer by one for each idle time slot observed on the channel.

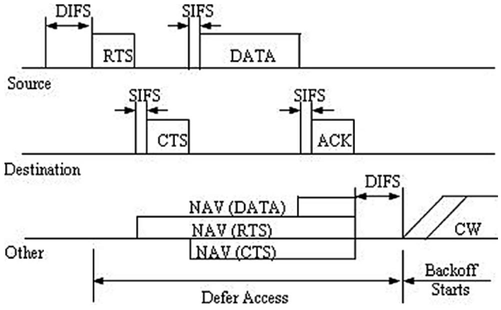


Fig. 19.2. RTS/CTS access mechanism in DCF

When the channel becomes busy due to other nodes' transmission ativities the node freezes its backoff timer until the channel is sensed idle for another DIFS. When the backoff timer reaches zero, the node begins to transmit. If the transmission is successful, the receiver sends back an acknowledgment (ACK) after an interval called the SIFS. Then, the transmitter resets its CW to CW_{min}. In case of collisions the transmitter fails to receive the ACK from its intended receiver within the specified period, it doubles its CW subject to maximum value CW_{max}, chooses a new backoff timer, and starts the above processes again.

In 802.11, DCF also provides a more efficient way of transmitting data frames that involve transmission of special short RTS and CTS frames prior to the transmission of actual data frame. As shown in Fig. 19.2, an RTS frame is transmitted by a node, which needs to transmit a packet. When the destination receives the RTS frame, it will transmit a CTS frame after SIFS interval immediately following the reception of the RTS frame. The source station is allowed to transmit its packet only if it receives the CTS correctly. Note that all the other stations are capable of updating their knowledge about other nodes' transmission duration by receiving a certain field in RTS, CTS, ACK, and packets transmission called network vector allocation (NAV). This helps to combat the *hidden terminal* problems. In fact, a node that is able to receive the CTS frames correctly can avoid collisions even when it is unable to sense the data transmissions directly from the source station. If a collision occurs with two or more RTS frames, much less bandwidth is wasted when compared with the situations where larger data frames are in collision, thus justifying the case for RTS, CTS mode of operation.

According to the protocol new users typically get the access before existing users who may collide with each other leading to fairness problems [3–13].

19.3 System analysis for the ideal standard case without backoff

Let p be the transmission probability of each node and M be the number of active stations. Assuming no backoff and each user tries to transmit randomly in each slot following the DIFS period, only one user tries his RTS which is then followed by CTS and a successful packet; the probability of successful transmission is thus given by the following

$$P_s = Mp(1-p)^{M-1} \tag{19.1}$$

The probability of an idle slot is

$$P_o = (1-p)^M \tag{19.2}$$

and probability of unsuccessful transmission RTS (collision) is

$$P_c = 1 - P_s - P_o \tag{19.3}$$

Let i be the number of idle periods (cycles) to success shown in Fig. 19.3 and j be the number of idle slots in each idle period lengths $\left(W_1, W_{2,}\right)$. So the efficiency (η_1) is given by Eq. (19.7).

It is easily seen that the average length of each idle period except the last one before packet success is given by

$$W_1 = W_2 = \ldots W_{I-1} = E(j) = \sum_{j=1}^{\infty} j\left(P_o\right)^j P_c$$

$$W_1 = W_2 = \ldots W_{I-1} = \frac{P_o P_c}{(1-P_o)^2} \text{ slots} \tag{19.4}$$

The last idle period has an average of

$$W_I = \frac{P_o P_s}{(1-P_o)^2} \text{ slots} \tag{19.5}$$

The average number of cycles is given by

$$I = \sum_{i=1}^{\infty} iP_i$$

$$I = \frac{(1-P_o)}{P_s} \tag{19.6}$$

All cycles leading to no success (RTS heard but no CTS) will each have a cost of $W_i + T_{RTS} + T_{DIFS} + T_{Slot} + T_{SIFS}$ seconds.

$$n_1 = \frac{T_{Payload}}{(W_1 + W_2 + W_{I-1})T_{Slot} + (I-1)\{T_{RTS} + T_{DIFS} + T_{SIFS}\} + T_{DIFS} + T_{RTS} + T_{CTS} + T_{ACK} + 3T_{SIFS} + T_{Payload} + W_1 T_{Slot}} \tag{19.7}$$

The number of collisions is $\bar{C} = I - 1$. Thus $\bar{C}$ and W_I are calculated from different values of *M* and *p* and stored in two **tables** (not shown for space consideration). So for particular values of *M* and *p* there is a particular value of $\bar{C}$ and W_I.

From Eq. (19.7) the efficiency η_1 can be calculated for different values of *M* and *p* as in Fig. 19.4. Table 19.1 depicts the probabilities at which the maximum efficiency occurs for different values of *M*.

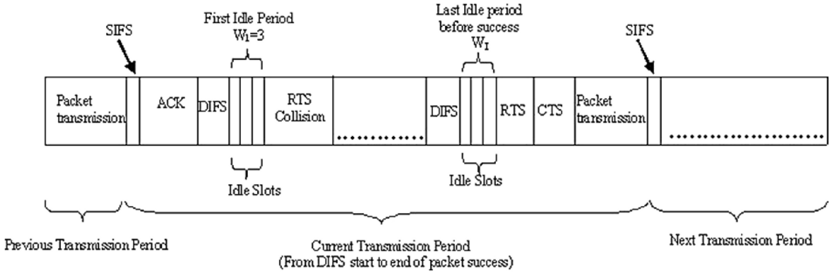


Fig. 19.3. Transmission activity on the wireless channel

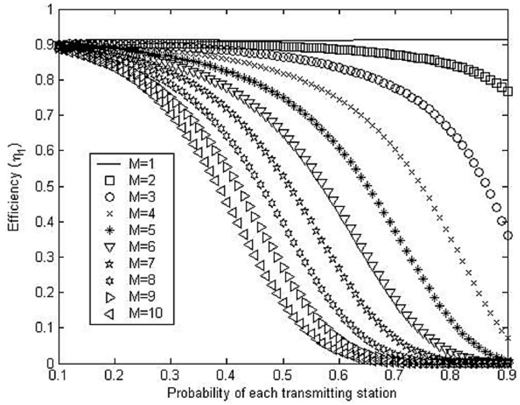


Fig. 19.4. Efficiencies for different probabilities and different number of stations

Table 19.1. Optimum efficiencies for different probabilities and different number of stations

No of stations	Probability	Optimum efficiency
1	0.9	0.914494552
2	0.25	0.903305479
3	0.15	0.901342654
4	0.11	0.900459291
5	0.1	0.89970777
6	0.1	0.898158644
7	0.1	0.895995929
8	0.1	0.893367296
9	0.1	0.890347837
10	0.1	0.886975732

19.4 Table-driven WLANs

In this new protocol, if the nodes sense that the channel is idle for an interval called DIFS (DCF interframe space), they try to send RTS of a packet with a probability *p* which is dependent on the traffic condition, i.e., the number and activities of the nodes as follows.

The users continuously monitor the channel in each idle slot following the DIFS idle period. If the previous slot is idle, it calls a uniform random generator (0,1). If the value of this generator is less than or equal to *p*, it tries to start its RTS transmission in the given next slot. If the value is larger than *p*, the users persist on listening and repeat RTS transmission trials as stated. However, if the channel is sensed busy the user defers his transmission till the next DIFS idle period heard.

The nodes measure the number of collisions $\bar{C} = I - 1$ and the length W_I of the last idle period sliding (by monitoring the channel over a large enough window), they can then use the tables formulated in Section 19.3 to obtain the corresponding p and *M*.

Users having a non-empty queue start by monitoring the channel for the first n transmission periods. This active user will average the length of the idle period preceding the correct packet transmission over n transmission periods, i.e., $\bar{W}_I$ and $\bar{C} = I - 1$, i.e., the average number of collision over the same period. Aided with these values the users obtain the operating values of p and *M* and use p to control their activities for the head of line packet in their queue. Active users continuously monitor the channel and use a sliding window technique to estimate W_I and *I* and hence obtain *M*, *p*. For example, the first sliding window averages W$_I$ and $\bar{C}$ of the first n transmission periods. The second window averages W_I and $\bar{C}$ of the $l = 2,3,...,n+1$ transmission periods, the third to (n+2) transmission periods. The sliding window averaging process reflects the changing traffic, so transmission activity of active users is dependent only on the current traffic and not on past history.

It is possible that the tables relating (M, p) to (W_I, I) yield more than one possibility for M, p for certain W_I, I measurement values from the sliding window. In this case, the users average the obtained values of *M* and use Table 19.1 to find the optimum p at this averaged value of *M*. This Table 19.1 is obtained from Fig. 19.4 in an evident manner. The operation of this table-driven technique is similar to the DCF standard IEEE protocol [4] except for using this optimized transmission probability p and discarding the timers and backoff windows. The active users just estimate *M*, *p* from the traffic conditions (by sensing the channel) in a sliding window fashion transmission, one period after another.

We note that both old and fresh users measure the traffic and adopt to same traffic condition fairly and obtain same p. However, having same p does not mean all users will repeatedly collide in the same slot, since feeding a random number generator with p yields different slot numbers to start transmitting the RTS each time it is called.

19.5 Buffer-adaptive WLANs

In the buffer-adaptive technique, each node's probability of transmission's trying is calculated based on the number of packets stored in the buffers. The probability of trying, p, of each node is $\frac{Bu}{Bu_{\max}}$, where Bu is the number of packets stored in the buffers at each node and $Bu_{\max}$ is the user buffer capacity. The buffer-adaptive technique is simple, it is similar to the IEEE 802.11 RTS/CTS, DCF mode except for the elimination of timers and exponential backoff. It is also similar to the table driven technique except for the elimination of table construction simply the users adjust their random transmission probability p following the standard DIFS period, continuously based on the queue size. This technique does need neither traffic measurement nor table establishments.

19.6 Simulation results

For numerical calculations the following parameters are used:
$T_{Payload}$=10msec; PHYheader=128bits; *ACK*=112bits+PHY header; *RTS*=160bits+PHY header; *CTS*=112bits+PHY header; *Channel bit rate*= 1 Mbits/s; Slot time (T_{Slot})= 50 µs; T_{SIFS}=28 µs; T_{DIFS} =128 µs.

In the table-driven technique, as per the standards, following the observance of each DIFS, users try to transmit with probability *p* obtained from the above table which is obtained from the traffic measurements. So all users with non-empty buffer try to transmit a packet with a probability obtained from the table. If two or more stations try to transmit at the same time, collisions occur. If no stations transmit (Fig. 19.3), the number of idle slots will increase. If one station is successful after a certain number of idle and collision period, the transmission period ends. As a result the total time for one successful packet transmission includes T_{DIFS}, T_{SIFS}, T_{RTS}, T_{CTS}, T_{Idle}, $T_{Payload}$. The efficiency is calculated at the end of the simulation at certain values of M, λ, p, i.e.,

$$\eta = \frac{T_{Payload} \times No\ of\ Transmissi\,on\ Periods\ in\ the\ whole\ simulation}{Time^{(n)}}$$

where $Time^{(n)}$ is the total simulation time which depends on T_{DIFS}, T_{SIFS}, T_{RTS}, T_{CTS}, T_{Slot}, $T_{Payload}$.

Initially $Time^{(n)} = T_{DIFS}$ and subsequently increased based on the user's activity, e.g.,

$$Time^{(n)} = Time^{(n)} + T_{Slot},\ \mathit{foreach\ idle\ slot\ \ period}$$

$$Time^{(n)} = Time^{(n)} + T_{RTS} + T_{DIFS} + T_{SIFS},\ \mathit{for\ \ each\ collision}$$

$$Time^{(n)} = Time^{(n)} + T_{RTS} + T_{CTS} + T_{DIFS} + T_{3SIFS} + T_{Payload},$$

$$\mathit{for\ each\ successful\ \ packet}$$

In the case of the buffer-adaptive technique, the simulation time is calculated in the same way as per user's activity above. However, no table is constructed and sliding windows do not apply.

In the table-driven technique described in Section 19.4 the active stations estimate the value of *M*. At certain traffic the curve shown in Fig. 19.5 is linear, which means the offered and the estimated values of the number of active stations are the same.

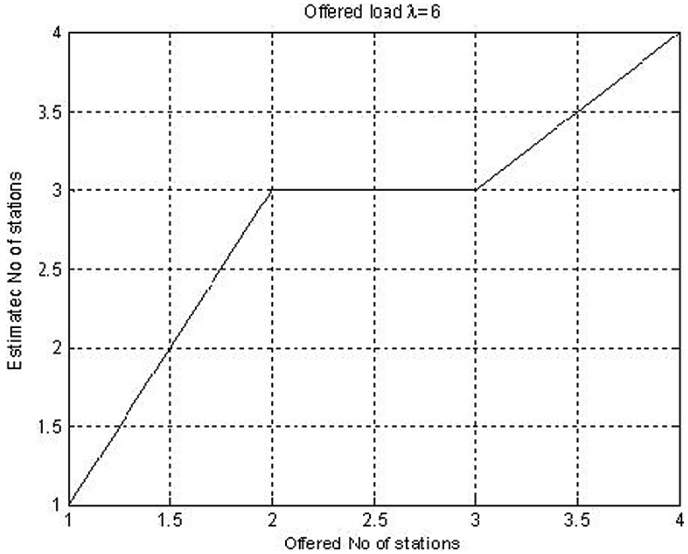


Fig. 19.5. Estimated *M* at a certain traffic

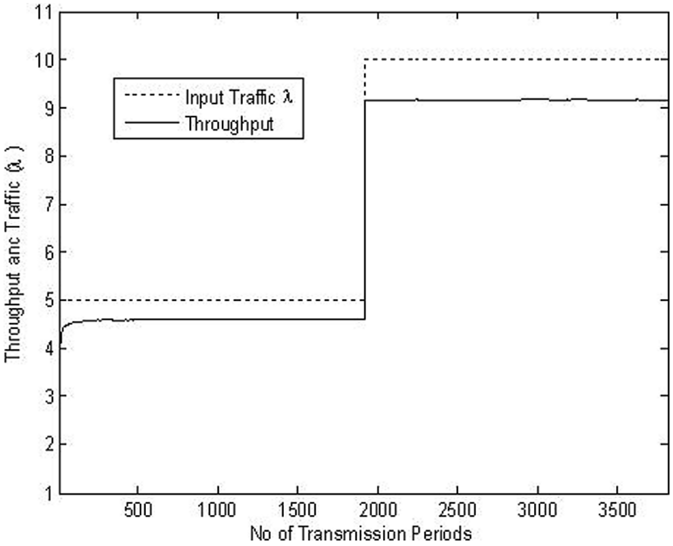


Fig. 19.6. Throughput and input traffic corresponding to the number of transmission periods

The table-driven technique can be considered as a load adaptive system. That means it has the capability to adapt to the input traffic as quickly as possible. Figure 19.6 shows a case where the input traffic suddenly increases from 5 packets/s to 10 packets/s. In this case the throughput $\left(\eta \times \mathit{Input\ traffic\ rate}(\lambda)\right)$ (Fig. 19.6) is shown to follow the offered traffic λ.

Figure 19.7 shows the efficiency curve for different offered loads for the table-driven technique for different window sizes. This shows that the efficiency rises and becomes saturated at higher values of the load. The window size has small effects on the efficiencies at different loads.

Figure 19.8 depicts the packet delay corresponding to different loads for the table-driven technique for different window sizes. The window sizes have little effect on the packet delay corresponding to different loads.

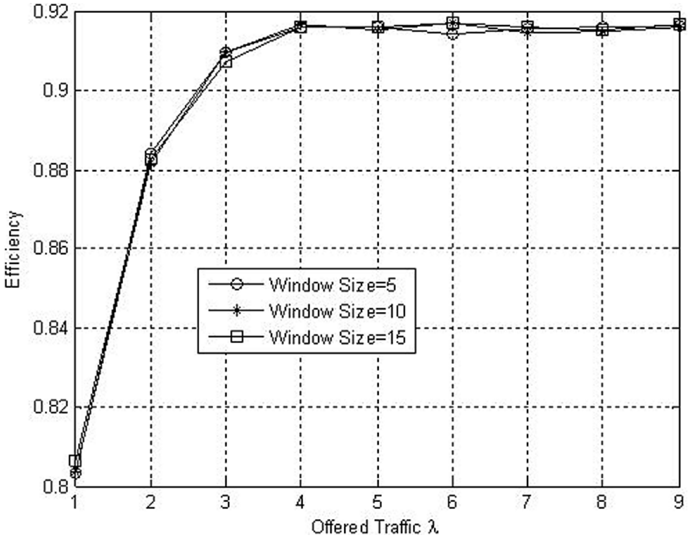


Fig. 19.7. Efficiency corresponding to different offered traffic using different window size

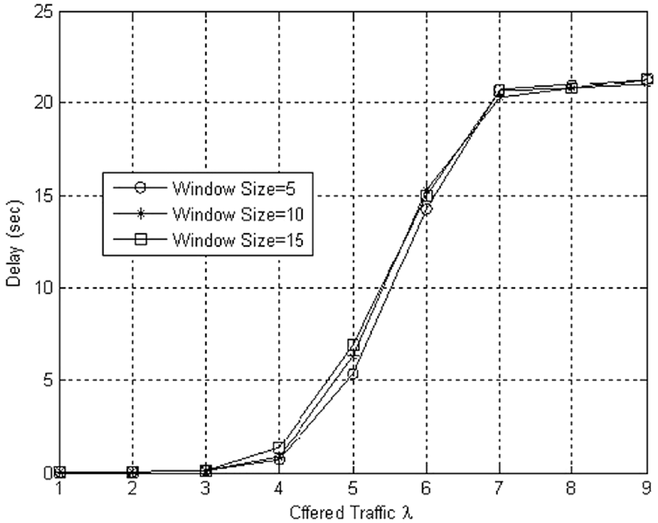


Fig. 19.8. Delay corresponding to different offered traffic using different window size

As the efficiency rises with the increased load, excess numbers of packets are left in the buffers. This results in a large packet delay at higher loads.

Figures 19.9 and 19.10 show the efficiency and the delay performance at different loads for the buffer-adaptive technique. From 2 packets/s to 4 packets/s, the efficiency of the buffer-adaptive technique and the table driven technique is more or less the same. Beyond four packets/s, in the buffer-adaptive technique the efficiency becomes very small, because all stations transmit their packets with higher probability which results in high amount of collision and no success. In these figures efficiency and delay are calculated for different buffer capacities, such as 30, 100, 300.

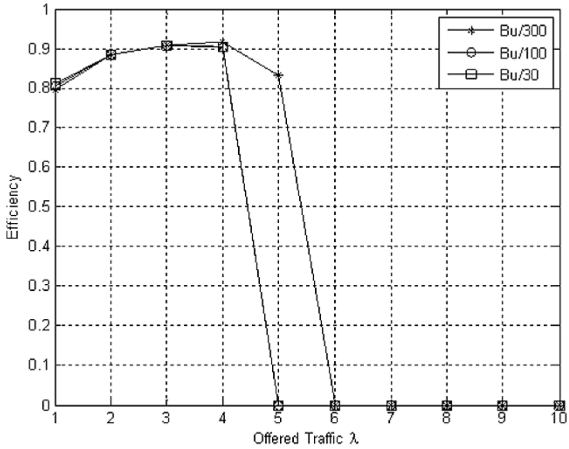


Fig. 19.9. Efficiency at different loads for buffer-adaptive system for different capacities

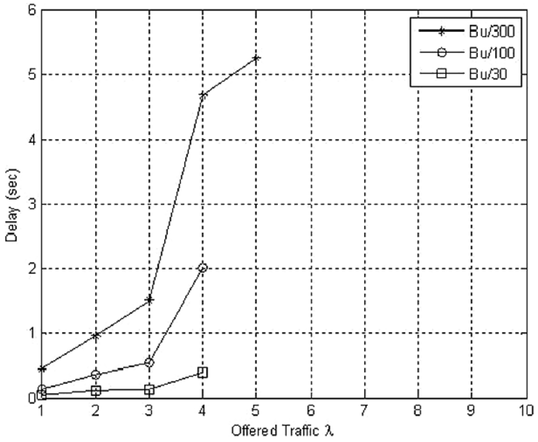


Fig. 19.10. Delay curves at different loads for buffer-adaptive system for different capacities

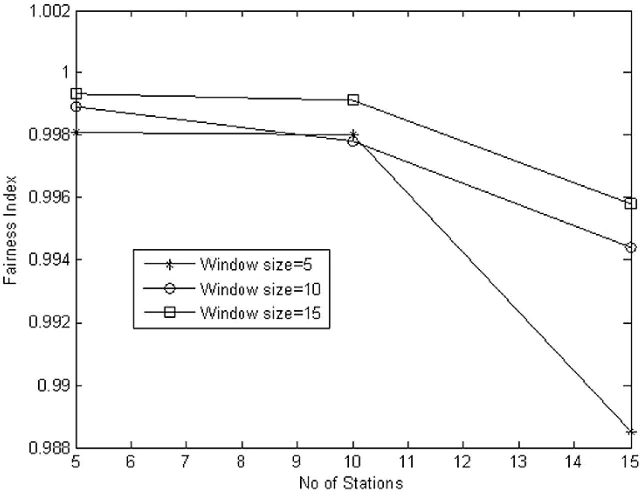


Fig. 19.11. Fairness index for the table-driven technique for different window sizes

Figure 19.10 shows that the delay performance degrades as the capacity of the buffer increases. However, the efficiency increases as the buffer capacity increases (Fig. 19.9).

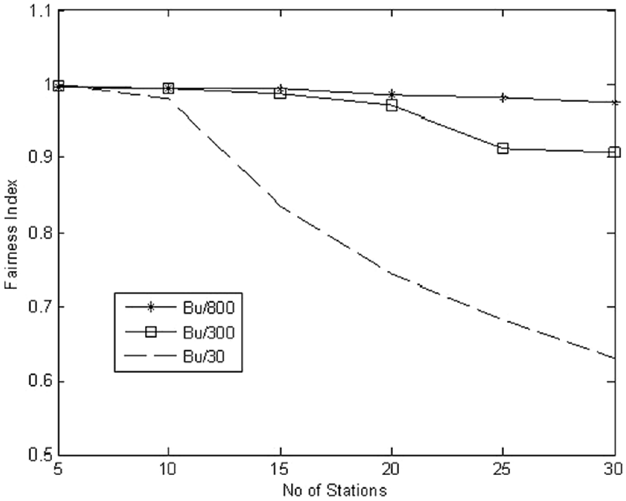


Fig. 19.12. Fairness index of the buffer-adaptive technique for different buffer capacities

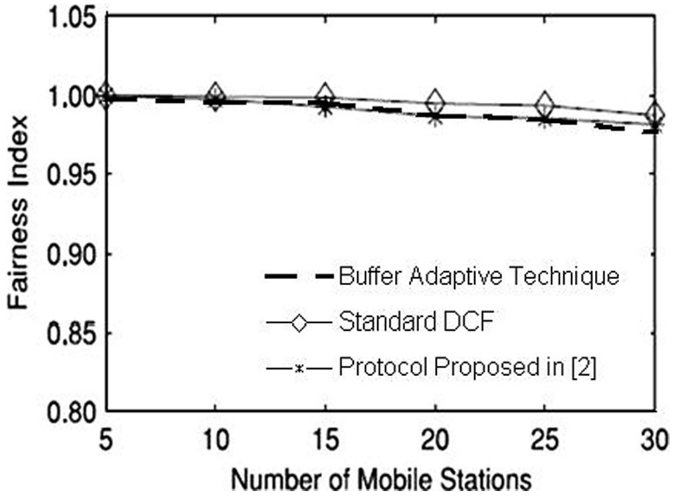


Fig. 19.13. Fairness index for different number of stations for different cases

Fairness is another important issue. To express this, we take the fairness index defined in [2] to measure the fair packet capacity allocation. That is,

$$FI = \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n\left(\sum_{i=1}^{n} x_i^2\right)},$$

where FI is the fairness index, n is the number of stations, x_i is the packets transmitted by the ith active station during the simulation time (current traffic in which the offered traffic λ is same for all stations).

Figure 19.11 shows that the fairness index decreases as the window size is decreased for the case of table-driven technique.

Figure 19.12 shows the fairness index of the buffer-adaptive technique. From this we can observe that the stations can be fairly operated when the buffer capacity is made high.

Figure 19.13 shows a comparison of the fairness index between the buffer-adaptive technique and the protocol proposed in [2]. We conclude that the buffer-adaptive technique yields the same fairness index as in [2] which uses modified exponential backoff.

Figure 19.14 shows a comparison between table-driven technique and buffer-adaptive technique along with the protocol proposed in [2]. It can be observed that for all the cases up to 20 active stations the performance is the same. Beyond that load, the fairness of the table-driven technique degrades.

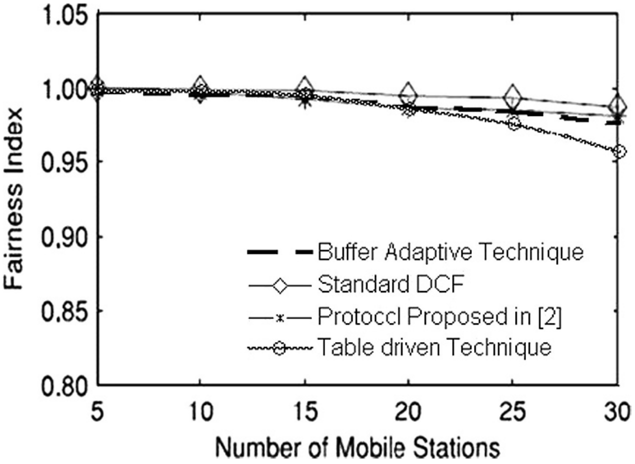


Fig. 19.14. Fairness index for different number of stations for different cases

Now let us introduce the delay variance for the two different new techniques. Delay variance is calculated by

$$DV = \frac{\sum_{i=1}^{n}\left(D_i - D_{average}\right)^2}{n},$$

where DV is the delay variance, D_i is the average packet delay of the ith station, and $D_{average}$ is the average packet delay of all stations.

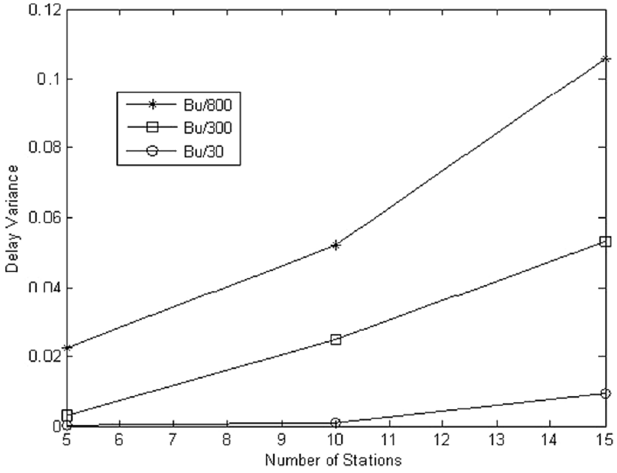


Fig. 19.15. Delay variance for the buffer-adaptive technique for different buffer capacities

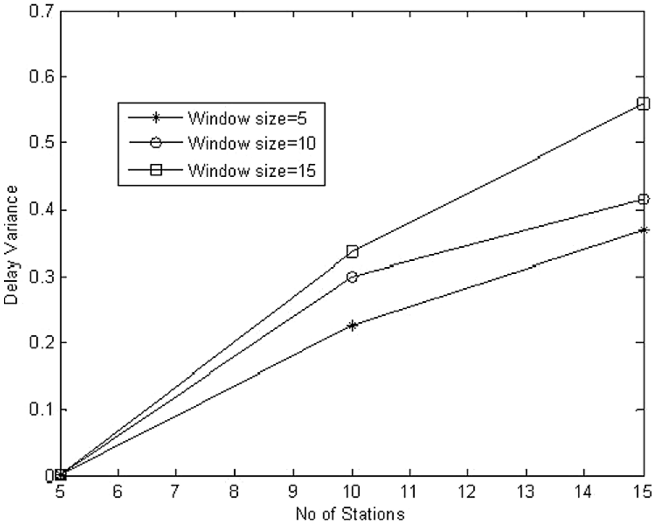


Fig. 19.16. Delay variance for the table-driven technique for different window sizes

Figure 19.15 shows the delay variance for the buffer-adaptive technique for different buffer capacities. It is observed that the delay variance increases in accordance with the number of stations as well as the buffer capacity.

Figure 19.16 shows the delay variance of the table-driven technique for different window sizes. The delay variance performance is worse than the case of buffer-adaptive technique.

19.7 Conclusion

In this chapter two new techniques (table-driven and buffer-adaptive) were presented, modeled, and compared.

Simulation results show that the table-driven technique performs well for faster load variations, whereas the buffer-adaptive technique does not. But the buffer-adaptive technique has FI (fairness index) and DV (delay variance) performances better than the table-driven technique making it suitable for real-time application (voice, video, etc). The buffer-adaptive technique does not scale well at higher loads as compared to the table-driven technique. The throughput and delay performances are better in the case of table-driven technique making it suitable for data application.

References

1. Xuejun T, Xiang C, Tetsuo I, Yuguang F (2005) Improving throughput and fairness in WLANs through dynamically optimizing backoff. IEICE Trans. Commun. E88-B(11):4328–4338
2. Liang Z, Yantai S, Oliver Y (2007) Performance improvement for 802.11 based wireless local Area networks. IEICE Trans. Commun. E90-B(4)
3. Bharghvan V (1998) Performance evaluation of algorithms for wireless medium access. IEEE International Computer Performance and Dependability Symposium IPDS'98, pp 86–95
4. Bianchi G (2000) Performance analysis of the IEEE 802.11 distributed coordination function. IEEE J. Sel. Areas Commun. 18(3):535–547
5. Tay YC, Chua KC (2001) A capacity analysis for the IEEE 802.11 MAC protocol. ACM/Baltzer Wireless Networks 7(2):159–171
6. Kim JH, Lee JK (1999) Performance of carrier sense multiple access with collision avoidance protocols in wireless LANs. Wirel. Pers. Commun. 11(2):161–183
7. Cali F, Conti M, Gregori E (2000) Dynamic tuning of the IEEE 802.11 protocol to achieve a theoretical throughput limit. IEEE/ACM Trans. Netw. 8(6):785–799
8. Kwon Y, Fang Y, Latchman H (2003) A novel MAC protocol with fast collision resolution for wireless LANs. IEEE INFOCOM'03
9. Weinmiller J, Woesner H, Ebert JP, Wolisz A (1996) Analyzing and tuning the distributed coordination function in the IEEE 802.11 DFWMAC draft standard. In: Proc. MASCOT, San Jose, CA
10. Wu H, Peng Y, Long K, Cheng S, Ma J (2002) Performance of reliable transport protocol over IEEE 802.11 wireless LAN: Analysis and enhancement. IEEE INFOCOM'02 2:599–607
11. Chhaya HS, Gupta S (1997) Performance modeling of asynchronous data transfer methods of IEEE 802.11 MAC protocol. ACM/Baltzer Wireless Networks 3:217–234
12. Bianchi G, Tinnirello I (2003) Kalman filter estimation of the number of competing terminals in an IEEE 802.11 network. IEEE INFOCOM'03 2:844–852
13. Eshaghi F. Elhakeem A. Shayan YR (2005) Performance Evaluation of Multihop Ad Hoc WLANs. IEEE Communications Magazine

Chapter 20

A new algorithm in blind source separation for high-dimensional data sets such as MEG data

Jalil Taghia[1], M. A. Doostari[1], Jalal Taghia[2]

[1] Department of Electrical Engineering, Shahed University, Khalije Fars Highway- 3319118651, Tehran, Iran, {Taghia, doostari}@shahed.ac.ir
[2] Department of Electrical and Computer Engineering, Shahid Beheshti University, Evin, 1983963113, Tehran, Iran, j.taghia@mail.sbu.ac.ir

Abstract. BSS is one of the well-known methods of signal processing. This method is based on recovering of original sources from observed mixtures without any further information about mixing system and original sources. In many applications, mixtures are combination of non-Gaussian and time-correlated components. MCOMBI algorithm is known as a method for separation of these kinds of sources. The performance and accuracy of this algorithm are noticeable but the high computational cost is one of the most significant limitations of MCOMBI algorithm, especially for high-dimensional data sets like high-density electroencephalographic (EEG) or magnetoencephalographic (MEG) recordings. In this chapter, we propose a new algorithm which uses combination of WASOBI and EFICA algorithms. In addition we use clustering method to decrease computational cost. In contrast with MCOMBI algorithm, the proposed algorithm decreases run time of separation and it has high accuracy close to MCOMBI algorithm. Thus, this algorithm is suitable for real high-dimensional data sets. In this chapter we use our algorithm for separation of artifacts in real MEG data.

Keywords. Statistical signal analysis, biomedical signal processing, blind source separation, independent component analysis, Non-Gaussianity, time-correlation, MEG data

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_20,

20.1 Introduction

Blind source separation (BSS) is one of the famous methods in signal processing. This method estimates original sources only from the observed mixtures and separating operation performed without any prior information about the mixing system. In this chapter we consider the most common BSS problem in which the sources are assumed to be independent and the mixing system is assumed to be linear and instantaneous. The BSS model can be expressed as follows:

$$x(t) = \sum_{j=1}^{d} a_j s_j(t) = As(t) \tag{20.1}$$

where $A = [a_1,...,a_d]$ is an unknown mixing matrix, $S(t) = [s_1(t),...,s_d(t)]^T$ are the original unobserved sources, and $X(t) = [x_1(t),...,x_d(t)]^T$ are the observed linear and instantaneous mixtures. The BSS problem consists in estimating a separating matrix $\widehat{W} \approx A^{-1} = W$ such that the mixing process A can be inverted and the sources S recovered: $\widehat{S} = \widehat{W}X = \widehat{W}AS \approx S.$

Different methods to solve the BSS problem usually differ in the statistics measuring the independence of the source signals and the estimators of those statistics. In fact the suitable choice for independence measure is more important than the selection of accuracy of the estimator. The optimal choice for independence measures depends on the generating model of the source signals. Non-Gaussianity and cross-correlation are two of the most common choices for measuring independence. When the original source signals are independent and identically distributed (i.i.d) processes with non-Gaussian distribution, we can perform separating operation by using maximizing measure of the estimated sources. Non-Gaussianity can be measured using marginal entropy for which several accurate estimators have been proposed in the BSS [1, 2]. EFICA [3] and FastICA [1] are best algorithms in the sense of speed and accuracy that are categorized in this group. On the other hand, the original sources can be identified by minimizing cross-correlation and maximizing auto-correlation in the estimated sources when the sources are time series with non-zero auto-correlation for time lags greater than zero. SOBI/TDSEP [4, 5] and WASOBI [6–8] are well-known algorithms in this group. In many real applications, mixtures are combination of non-Gaussian and time-correlated sources. Thus only some of the sources (not all of them) can be separated by applying algorithm that use only one of the independence measures. Probably a more accurate approach is to use algorithms based on combination of non-Gaussianity and cross-correlation measures. Methods based on this idea are EFWS [9], COMBI [9], and MCOMBI [10]. A practical limitation of all these combination approaches is that their computational cost is unaffordable for high-dimensional mixtures like the ones found in high-density electroencephalography (EEG) and magnetoencephalography (MEG).

In this chapter, we propose a new algorithm that uses a combination of WASOBI and EFICA algorithms (like MCOMBI algorithm). In addition we use clustering method to decrease computational cost. In contrast with MCOMBI algorithm, the proposed algorithm decreased run time of separation with high accuracy close to MCOMBI algorithm. We show that applying of this algorithm is suitable for real high-dimensional data sets. In this chapter we use this algorithm for separation of artifacts in real MEG data.

20.2 Multidimensional independent components

Standard BSS assumes that the 1-D unknown sources in Eq. (20.1) are mutually independent according to the independency contrast used. A generalization of this principle assumes that not all the d sources are mutually independent but they form M higher dimensional independent components [11, 12]. Let d_l denote the dimensionality of the lth multidimensional component that groups together the 1-D source signals with indexes $l_1,\ldots,l_{d_l}$. Then, the lth multidimensional component is given by $S_l = [s_{l_1},\ldots,s_{l_{d_l}}]^T$, where $l = 1,\ldots,M$ and $d_1 + d_2 + \cdots + d_M = d$. Therefore, we can rewrite the sources data matrix S in Eq. (20.1) as $S = [s_1,\ldots,s_d]^T = Q[S_1,\ldots,S_M]^T$ where Q is a permutation matrix. Using the above notation and dropping matrix Q under the permutation indeterminacy of ICA, we can reformulate Eq. (20.1) as follows:

$$S = WX = [W_1X,\ldots,W_MX]^T = [S_1,\ldots,S_M]^T \tag{20.2}$$

The goal of multidimensional BSS is to estimate the sub-matrices $\{W_l\}_{l=1,\ldots,M}$ each of which is of dimension $d_l \times d$. Since the subcomponents of a multidimensional independent component are arbitrarily mixed, we can recover $\{W_l\}_{l=1,\ldots,M}$ only up to an invertible matrix factor [12]. A multidimensional component according to certain independency contrast (e.g., non-Gaussianity) might be separable into 1-D components using an alternative independency measure (e.g., cross-correlations). This suggests a procedure for combining complementary independency criteria [10]:

1. Try BSS using certain independency criterion.
2. Detect the presence of multidimensional components in the source signals estimated in step 1.
3. Try BSS using an alternative independency contrast in each multidimensional component found in step 2.

This is the basic idea underlying proposed algorithm which combines the complementary strengths of the Non-Gaussianity criterion of EFICA and the criterion based on cross-correlations of WASOBI.

20.3 Detection of multidimensional independent components

A common way of evaluating the accuracy of the separation produced by any BSS algorithm is the matrix of interference-to-signal ratios (ISR matrix). Element wise, the ISR matrix is defined as $ISR_{kl} = G^2{}_{kl} / G^2{}_{kk}$ where $G = \widehat{W}A$. $\widehat{W}$ is the estimated separating matrix and A is the true mixing matrix. ISR_{kl} measures the level of residual interference between the kth and lth estimated components. The total ISR of the kth estimated source is defined as follows:

$$isr_k = \sum_{l=1, l \neq k}^{d} ISR_{kl}. \tag{20.3}$$

EFICA and WASOBI share the rare feature of allowing the estimation of the obtained ISR matrix through simple empirical estimate of $E[ISR]$ using the estimated sources $\widehat{S}$. This means that EFICA and WASOBI permit us to estimate $\widehat{ISR} \approx E[ISR]$. It has been shown that the estimations $\widehat{ISR}$ obtained by WASOBI and EFICA are quite accurate even when the respective assumptions about the sources are only partially fulfilled [10]. The information provided by $\widehat{ISR}$ is crucial for detecting the presence of multidimensional components within the estimated sources which is the reason for us to choose EFICA and WASOBI in our combined BSS method.

If the ISR matrix is known, or if it can be estimated, we can easily assess the presence of multidimensional independent components by grouping together components with high mutual interference. This is done by defining a symmetric distance measure between two estimated components as follows:

$$D(\hat{s}_k, \hat{s}_l) = D_{kl} = \frac{1}{ISR_{kl} + ISR_{lk}} \geq 0 \qquad \forall\ l \neq k \tag{20.4}$$
$$D_{kk} = 0 \qquad \forall\ k$$

Using the distance metric D, we cluster together the estimated components whose distance from each other is small. For this task we use agglomerative hierarchical clustering [13] with single linkage. By single linkage we mean that the distance between clusters of components is defined as the distance between the closest pair of components. The output of this clustering algorithm is a set of $i = 1,...,d$ possible partition levels of the estimated sources. At each particular level the method joins together the two clusters from the previous level which are closest in distance. Therefore, in level $i = 1$ each source forms a cluster whereas in level $i = d$ all the sources belong to the same cluster. For assessing the goodness-of-fit of the $i=2,...,d-1$ partition levels, we propose using the validity index $I_i = D_i^{\text{int}era} / D_i^{\text{int}er}$ where $D_i^{\text{int}era}$ and $D_i^{\text{int}er}$ roughly measure, respecttively, the average intra-cluster and inter-cluster distances. They are defined, for $1 < i < d$, as follows:

$$D_i^{\text{int}era} = \frac{\sum_{j=1, card(\Gamma_{i,j})>1}^{d-i+1} Card(\Gamma_{i,j})(Card(\Gamma_{i,j}) - 1)/2}{\sum_{j=1, card(\Gamma_{i,j})>1}^{d-i+1} \sum_{k\in\Gamma_{i,j}, l\in\Gamma_{i,j}} ISR_{kl}} \tag{20.5}$$

$$D_i^{\text{int}er} = \frac{\sum_{j=1}^{d-i+1} Card(\Gamma_{i,j})(d - Card(\Gamma_{i,j}))}{\sum_{j=1}^{d-i+1} \sum_{k\in\Gamma_{i,j}, l\notin\Gamma_{i,j}} ISR_{kl}}$$

where $\Gamma_{i,j}$ is the set of indexes of the sources belonging to the j th cluster at the ith partition level and $Card(\Gamma_{i,j})$ determines the number of elements in $\Gamma_{i,j}$. We also define $I_1 = 1/ISR_{\max}$ where $ISR_{\max}$ is the maximum entry in the ISR matrix. We set $I_d = 10$. Finally we choose the best cluster partition to be that one corresponding to the maximum of all local maxima of the cluster validity index I. By setting $I_d = 10$ we consider that the separation failed completely (there is just one d-dimensional cluster) if $D_i^{\text{inter}} < 10$, $D_i^{\text{intera}} \forall i = 2,...,d-1$. The definition $I_1 = 1/ISR_{\max}$ means that the estimated sources will be considered to be 1-D (perfect separation) if $ISR_{\max} < \min_{i>2}(1/I_i)$. Therefore, since $I_d = 10$, we require the maximum ISR between two 1-D components to be in any case below −10 dB. In order to ease the explanation of proposed algorithm in the next section we will use the following MATLAB notation to refer to the hierarchical clustering algorithm described in this section: $[i, I] = hclus(ISR)$ where the input parameter is the estimated ISR matrix, the first output parameter is the selected partition level and the second output parameter is a $1 \times (d-i+1)$ cell array such that $I\{k\}$ is a vector containing the indexes of the sources belonging to the kth cluster.

20.4 Proposed algorithm

The proposed algorithm is described using MATLAB notation as below:

```
function [B] = proposed-fun (X,ARorder)
[d, L] = size(X) ;
[B, ISRwa] = WASOBI (X,ARorder) ;
[iwa, Iwa] = hclus(ISRwa) ;
if iwa = = 1 , return; end
for i = 1: (d-iwa+1) ,
    if length (Iwa{i}) = = 1, continue; end
    index = Iwa{i}; di = length(index) ;
    [Bef, ISRef] = EFICA (B(index, :)*X) ;
    [ief, Ief] = hclus(ISRef) ;
    if (ief < di) || ...
      (min(sum(ISR(index,index),2)) > ...
       min (sum(ISRef,2))),
          B(index,:) = Bef*B(index,:) ;
    end
end
```

This algorithm starts by applying WASOBI on the input data. The reason for using WASOBI first instead of EFICA is that the former is considerably faster than the latter for high-dimensional mixtures, which is the main application of our algorithm. Subsequently, EFICA is applied on each multidimensional component of sources found in the output of WASOBI. Finally, we decide whether EFICA was able to improve the separation of the sources within the cluster or not. In our implementation of the algorithm we include a third step (not shown in the MATLAB code above) that consists in running WASOBI again on the cluster of unresolved components in the output of EFICA (if such a cluster exists). This last step is helpful only in the rare cases when, in the first run of WASOBI, we were not able to detect the correct clusters. If EFICA was able to separate some non-Gaussian sources we expect the accuracy of WASOBI to improve by applying it only to the cluster of Gaussian components that was not correctly separated by EFICA. WASOBI requires the user to specify the order of the AR model that best fits the unobserved sources. However, the performance is not critically dependent on this parameter and it is enough to select an order high enough to model appropriately the source signals.

20.5 Separation via proposed algorithm

In this section we use proposed algorithm for separation of artifacts in MEG data. Moreover we compare speed and accuracy of this algorithm with MCOMBI algorithm.

20.5.1 Introducing used MEG data

Magnetoencephalography (MEG) is a noninvasive technique by which the activity or the cortical neurons can be measured with very good temporal resolution and moderate spatial resolution. When using a MEG record, as a research or clinical tool, the investigator may face a problem of extracting the essential features of the neuromagnetic signals in the presence of artifacts. The amplitude of the disturbances may be higher than that of the brain signals, and the artifacts may resemble pathological signals in shape.

In this chapter we employ our algorithm to separate brain activity from artifacts. The approach is based on the assumption that the brain activity and the artifacts (e.g., eye movements or blinks, or sensor malfunctions) are anatomically and physiologically separate processes and this separation is reflected in the statistical independence between the magnetic signals generated by those processes. The MEG signals were recorded in a magnetically shielded room with a 122-channel whole scalp Neuromag-122 neuromagnetometer. This device collects data at 61 locations over the scalp, using orthogonal double-loop pickup coils that couple strongly to a local source just underneath. The test person was asked to blink and make horizontal saccades, in order to produce typical ocular (eye) artifacts. Moreover, to produce myographic (muscle) artifacts, the subject was asked to bite his teeth. Yet another artifact was created by placing a digital watch 1 m away from the helmet into the shielded room. In Fig. 20.1 we present a subset of nine observed MEG signals from the frontal, temporal, and occipital areas.

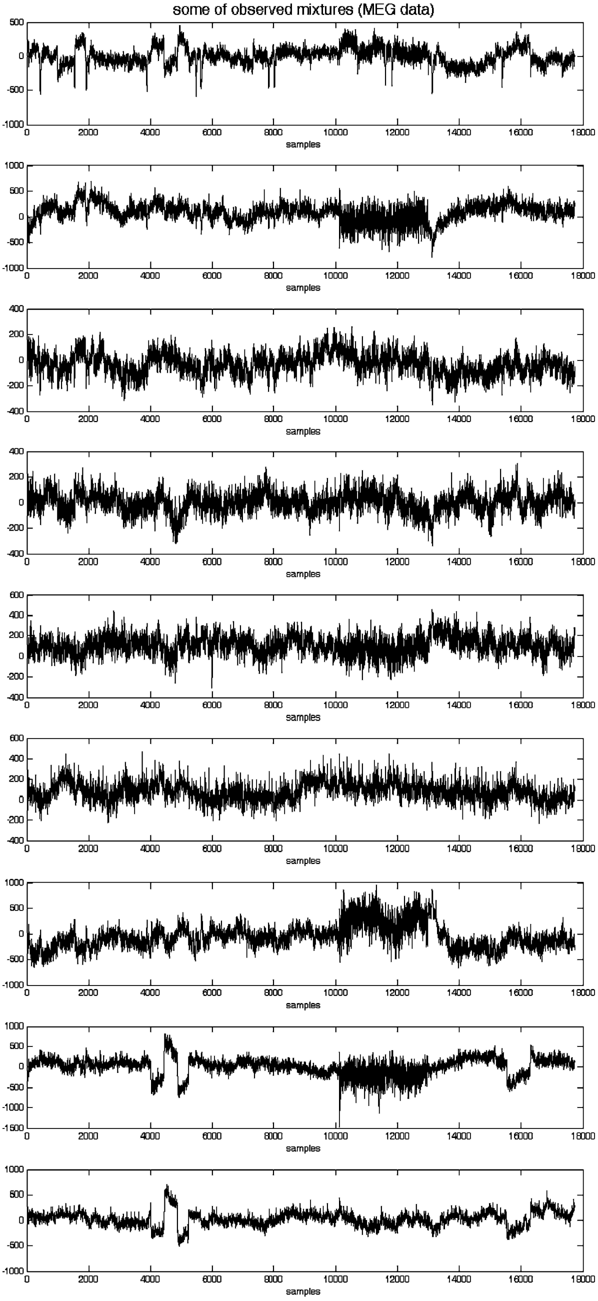


Fig. 20.1. Nine observed signals from MEG data

20.5.2 Experimental results

We applied proposed algorithm on MEG data (introduced in the last section) to separate artifacts from it. Figure 20.2 shows eight estimated independent components (ICs) that were found from the recorded data after applying proposed algorithm. The first two ICs (i.e., IC_1, IC_2) are clearly due to the muscular activity originating from the biting. IC_3 and IC_4 show the horizontal eye movements and the eye blinks, respectively. IC_5 represents the cardiac artifact that is very clearly extracted. Sixth independent component (i.e., IC_6) is due to breathing. To find the remaining artifacts, the data were high-pass filtered, with cutoff frequency at 1 Hz. Next, the independent component IC_7 was found. It shows clearly the artifact originated at the digital watch, near the magnetometer. The last independent component IC_8 is related to a sensor presenting higher RMS (root mean squared) noise than the others. To evaluate the overall separation performance of proposed algorithm and MCOMBI, we used the average of the ISR obtained for the individual sources, i.e.,

$$ISR_{avg} = \frac{1}{d} \sum_{k=1}^{d} isr_k \tag{20.6}$$

where d indicates number of multidimensional components and isr_k can be obtained from Eq. (20.3). In Fig. 20.3 we show the average signal-to-Interference Ratio (SIR) obtained for different number of data samples of the sources, to compare accuracy of two algorithms. In this figure, dashed line (--) and solid line present SIR_{avg} which is computed from MCOMBI and proposed algorithm, respectively. According to Fig. 20.3, the proposed algorithm is almost as accurate as MCOMBI algorithm.

On the other hand, the clustering method is used to decrease computational cost and running time of our algorithm with respect to MCOMBI algorithm. For comparing the speed of two algorithms, we have evaluated running time of them and presented results in Table 20.1. It can be understood from this table that the computation time of applying our algorithm on MEG data is clearly smaller than the computation time of applying MCOMBI. The major advantage of our algorithm is the possibility of using it with very high-dimensional data sets. It is noticeable that the obtained running time is an average of ten times iteration of two algorithms individually, and the implementation of two algorithms is performed using MATLAB v7.0.4 software in computer with below system properties, Intel Pentium 4 CPU 1.70 GHz/512 MB of RAM

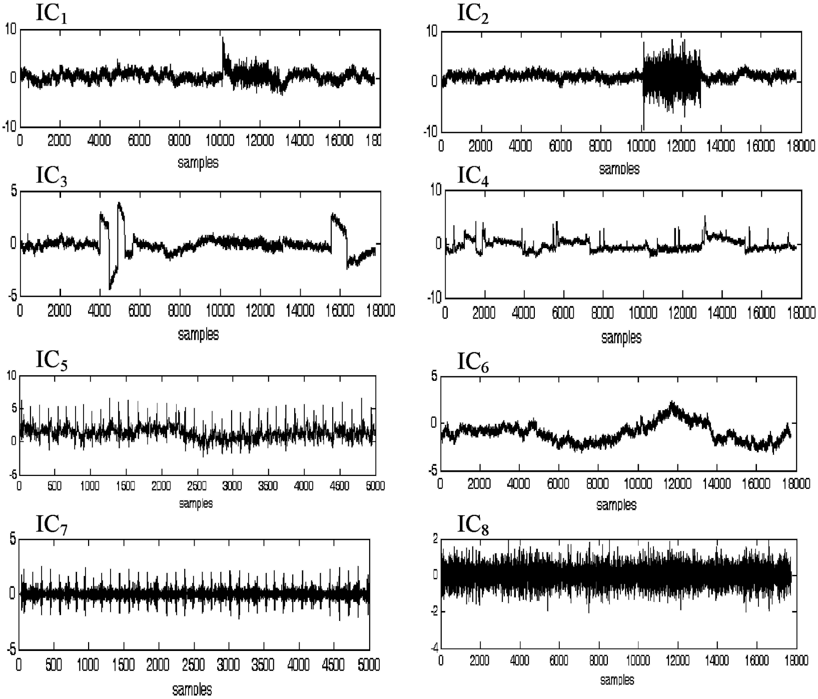


Fig. 20.2. Independent components (artifacts) estimated using proposed algorithm

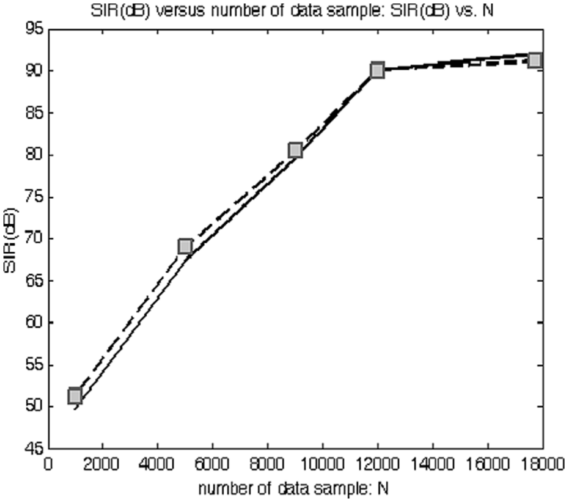


Fig. 20.3. Average signal-to-interference ratio (SIR_{avg}) obtained for different number of data samples of the sources. *Dashed line* (--) and *solid line* present SIR_{avg} that was computed from MCOMBI and proposed algorithm, respectively

Table 20.1. Running time of MCOMBI and proposed algorithms

Type of algorithm	Running time in seconds
MCOMBI	330
Proposed algorithm	42

20.6 Conclusions

We proposed a new BSS algorithm that is a combination of WASOBI and EFICA algorithms. This algorithm simultaneously separates non-Gaussian and time-correlated sources. Moreover we used clustering method for decreasing computational cost and running time. We applied our algorithm and MCOMBI algorithm on the real MEG data with high-dimensional data sets in order to separate artifacts from it. Furthermore, we demonstrated that this algorithm is almost as accurate as the MCOMBI algorithm. Moreover, because of using clustering method, our algorithm has lower computational cost with respect to MCOMBI. Thus, the proposed algorithm is suitable for the high-dimensional data sets.

Acknowledgment. We appreciate Doctor M. Kakooei and Pastor Hospital for the MEG data used as examples of a real biological environment.

References

1. Hyvärinen A (1999) Fast and robust fixed point algorithms for Independent Component Analysis. IEEE Transactions on Neural Networks, 10(3) pp 626–634
2. Cardoso JF (1999) High-order contrasts for Independent Component Analysis. Neural Computation, pp 157–192
3. Koldovský Z, Tichavský P, Oja E (2006) Efficient variant of algorithm Fastica for independent component analysis attaining the cramerrao lower bound. IEEE Transactions on Neural Networks, pp 1265–1277
4. Belouchrani A, Meraim KA, Cardoso JF, Moulines E (1997) A blind source separation technique based on second order statistics. IEEE Transactions on Signal Processing, pp 434–444
5. Ziehe A, Müller KR (1998) TDSEP-an efficient algorithm for blind separation using time structure. In: Proc. ICANN, pp 675–680
6. Yeredor A (2000) Blind separation of Gaussian sources via second-order statistics with asymptotically optimal weighting. IEEE Signal Processing Letters, pp 197–200

7. Doron E, Yeredor A (2004) Asymptotically optimal blind separation of parametric Gaussian sources. In: Proc. ICA, Granada, Spain
8. Tichavský P, Doron E, Yeredor A (2006) A computationally affordable implementation of an asymptotically optimal BSS algorithm for AR sources. In: Proc. EUSIPCO, Florence, Italy
9. Tichavsk´y P, Koldovsk´y Z, Doron E, Yeredor A, G´omezHerrero G (2006) Blind signal separation by combining two ICA algorithms: HOS-based EFICA and time structure-based WASOBI. In: Proc. EUSIPCO, Florence, Italy
10. Tichavsk´y P, Koldovsk´y Z, Yeredor A, G´omez-Herrero G, Doron E (2007) A hybrid technique for blind separation of Non-Gaussian and time-correlated sources using a multi-component approach. IEEE Transactions on Neural Networks
11. Lathauwer LD, Callaerts D, Moor BD, Vandewalle J (1995) Fetal electrocardiogram extraction by source subspace separation. In: Proc. IEEE Signal Processing Workshop on Higher-Order Statistics, Girona, Spain, pp 134–138
12. Cardoso JF (1998) Multidimensional independent component analysis. In: Proc. ICASSP, Seattle, WA
13. Winter S, Sawada H, Araki S, Makino S (2004) Hierarchical clustering applied to overcomplete BSS for convolutive mixtures. Workshop on Statistical and Perceptual Audio Processing SAPA, Korea

Chapter 21

Visible light source temperature estimation using digital camera photography

Anagha M. Panditrao[1], Priti P. Rege[2]

[1] Cummins College of Engineering for Women, panditraog@yahoo.co.uk
[2] College of Engineering, Pune, INDIA, ppr@extc.coep.org.in

Abstract. Light source temperature measurement has been an important aspect in many industrial applications. The aim of this study is to determine the temperature of different sources like flames, incandescent lamps. The sensor installation is difficult in these sources. Imaging these sources to correlate with the temperature is attempted. A digital still camera is used to capture the source images. The optical characteristics of CCD sensor are not mentioned in the camera specifications. It is desired that the device bandwidth should be greater than the source bandwidth. Hence, the optical characteristics of the camera are also obtained. As incandescent lamp is the most familiar light source, its photographs are taken. The images are taken at known excitation voltages in a dark room with camera settings unaltered. Filament temperature is a function of intensity. Intensity difference of two consecutive images is found out by image subtraction. Unwanted part of image, like reflection from the shell, is removed by thresholding and segmentation. The colour temperature of the filament can be found out with reference to black body radiations. If these values are compared with standard temperatures, source temperature values can be predicted.

Keywords. Non-contact temperature measurement, Imaging, Digital camera photography

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_21,

21.1 Introduction

This study deals with the determination of temperature in sources like furnace and incandescent lamp. The installation of sensor in above-mentioned sources is difficult. If the image of these sources is acquired by some means and post-processing is carried out, it is possible to predict the temperature. Most of the light sources emit light as a function of temperature. The light emitted by an incandescent source is a mixture of light with different wavelengths. The light from a black body is a mixture of light with a continuous range of wavelengths. An incandescent lamp is very nearly a black body radiator. This is the most commonly used light source. The distribution of power in the wavelengths it produces can be described by the temperature of a black body radiator whose light would appear to the human eye to be of the same colour [1]. The development of digital cameras has opened up possibilities for powerful new diagnostic techniques employing two-dimensional imaging. Digital still cameras (DSCs) have gained significant popularity in recent years [2]. Digital photography is the easily available fast technique and requires no other sensor. It is also possible to acquire the image to get total temperature distribution. Qualitative visualization for the temperature measurement is done using digital camera [3]. A Sony-make digital still camera DSC-S60 is used to acquire images. The optical characteristics of CCD sensor are not mentioned in the camera specifications. The proposed work is to obtain the characteristics of the digital camera and to obtain filament temperature of incandescent lamp by image processing.

21.2 Camera characteristics

In the proposed work a Sony-make digital still camera DSC-S60 is used to capture the images of various sources. As the optical characteristics are not mentioned in the specifications, different sources are used to cover the wide bandwidth ranging from IR to UV. It is desired that the device bandwidth should be greater than the source bandwidth. To verify if the camera is having adequate bandwidth, it is necessary to plot the camera characteristics.

21.2.1 Experimental setup

While designing the setup it is ensured that the effect of stray light is minimized and reflection of light is avoided. We have used following sources for our experimentation:

- 5 W, 12 V Lumina-make tungsten filament lamp
- LED: Colour (R, G, B), UV and IR

Enclosure length is adjusted so as to match camera minimum distance specifications. The setup photographs are shown in Fig. 21.1.

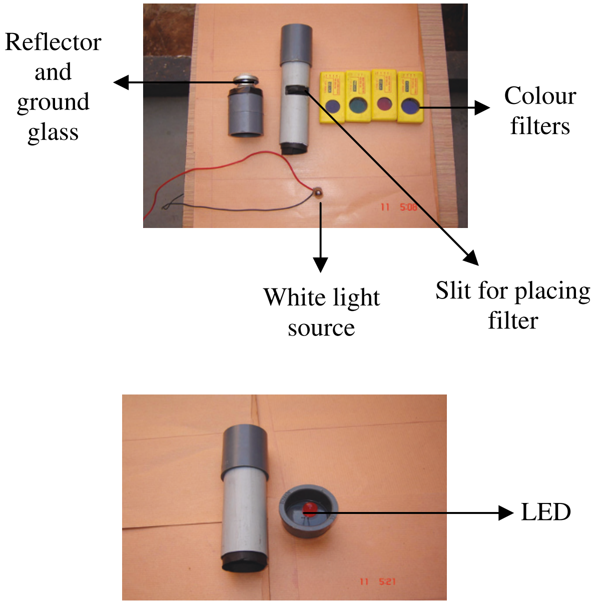
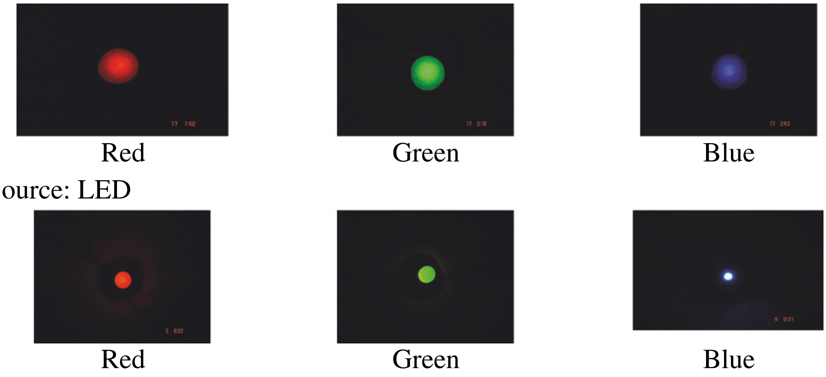


Fig. 21.1. Setup photographs


Red Green Blue

Source: LED

Red Green Blue

Fig. 21.2. Filter and LED images

Photographs of colour filters (red, green and blue) are taken where the source intensity is changed without affecting its white light nature. Similar technique is applied for LED array (red, green and blue).

The images obtained are shown in Fig. 21.2

21.2.2 Images obtained

Source: Filter

21.2.3 Processing

The colour values (red, green and blue) of the images acquired are separated as shown in Fig. 21.2. The values are compared with standard digital values. Red coloured images are compared with (255, 0, 0), green and blue are compared with (0, 255, 0) and (0, 0, 255), respectively. The processing is carried out using MATLAB software.

After processing the images and comparing the colour values with the standard values, it is observed that the obtained values are fairly matching with the standard values for R, G and B.

21.3 Source image acquisition

After obtaining the characteristics of camera, images of lamp filament are taken in a dark room with known excitation voltages. Figure 21.3 shows a filament image at a typical excitation voltage. In this photograph, section (a) is the filament image and section (b) is filament reflection from shell.

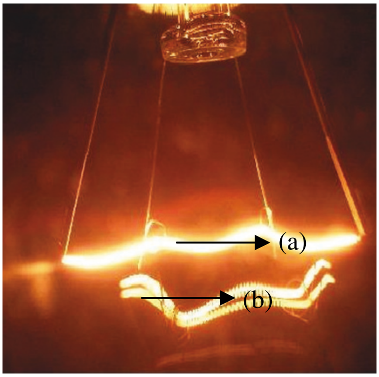


Fig. 21.3. Incandescent lamp filament image

The block schematic of the proposed system is given in Fig. 21.4.

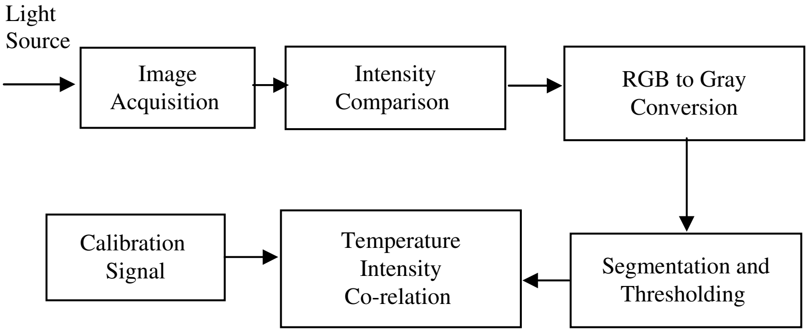


Fig. 21.4. Block schematic of the proposed non-contact source temperature measurement system

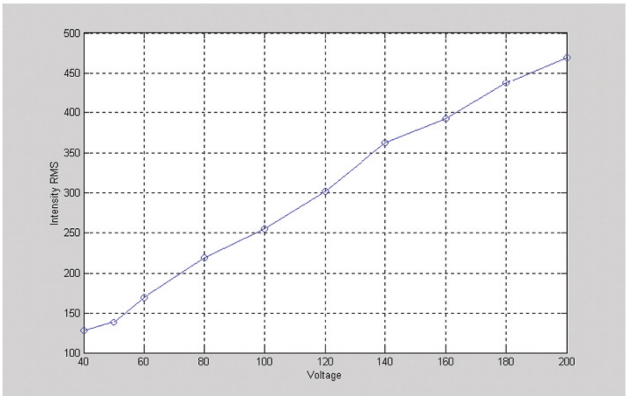


Fig. 21.5. Graph of intensity RMS value vs excitation voltage

Incandescent and fluorescent lights are two most common sources of artificial light. Incandescent light is created by passing an electric current through a resistance tungsten filament [3]. Characteristics of light can be described by the temperature. An incandescent lamp is connected to a dimmer. As the dimmer is turned up, the voltage increases and the lamp's filament becomes warmer and warmer until it begins to glow cherry red. As the voltage continue to increase, the filament gets hotter and hotter, glows more brightly and becomes less and less red [1]. The filament images of a 40 W Philips-make lamp are captured on digital camera. Incandescent lamps emit light solely because of their temperature. Tungsten wire temperature is a function of flow of electric current [4]. To avoid ambient light effect, the images are taken in a dark room. Images acquired using digital camera are used in further processing. Camera settings are not changed till all images are captured.

The graph of RMS value of intensity vs the excitation voltage is shown in Fig. 21.5. The intensity varies linearly with excitation voltage. The images obtained are converted to grey for further processing. Unwanted parts are removed using segmentation algorithm. The change in colour or intensity can be co-related with the filament temperature [5]. The temperature values can be compared with standard temperatures.

21.4 Image processing

The images acquired using a digital camera of known resolution are having natural colour components red, green and blue [6]. The intensities of two consecutive images are compared to view the change with respect to excitation voltage. Figure 21.6 shows this intensity comparison of images at 100 and 120 V.

These RGB components are converted to grey scale for further processing. To find out the intensity difference, images are converted to matrix form. The intensity component of every image is separated [6]. By subtracting the intensity matrices of two consecutive grey images intensity difference is obtained.

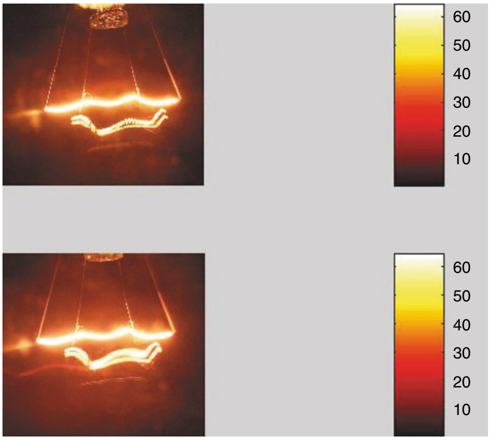


Fig. 21.6. Intensity comparison of two consecutive images

Figure 21.7 shows the difference between two images. Images (a) and (b) are the images at excitation voltage 200 and 230 V, respectively. Image (c) is the intensity difference between above-mentioned images.

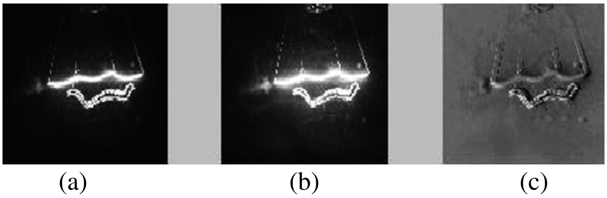


Fig. 21.7. Intensity difference between two images

In the images under study, along with the filament image, shell reflection is also seen. Hence, pre-processing of the images is required. In the analysis of area of interest in image, it is essential to distinguish between the object of interest, i.e. filament from the rest. Most commonly used techniques for segmenting foreground from the background are thresholding and edge detection. No segmentation technique is perfect and there is no universally applicable technique that works for all images [8]. Because of its intuitive properties and simplicity of implementation, image thresholding enjoys a central role in this application.

The proposed work is based on histogram-based thresholding. The brightness histogram of an image at typical excitation is shown in Fig. 21.8. In this histogram, *X*-axis represents the intensity of the pixels, whereas *Y*-axis represents the number of pixels having same intensity values.

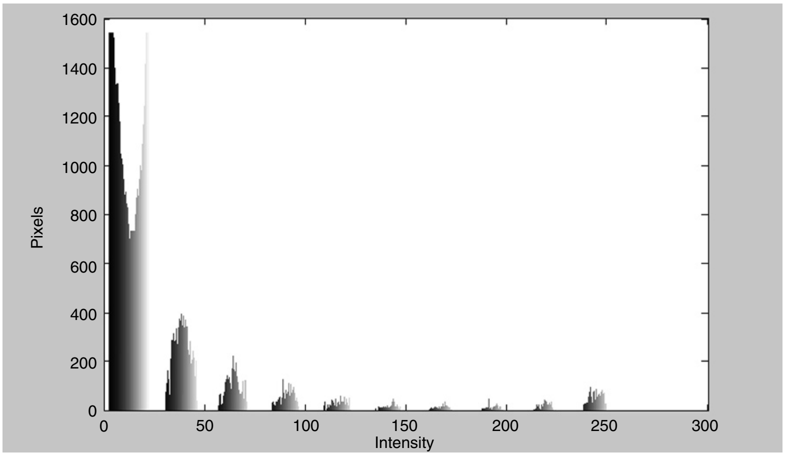


Fig. 21.8. Intensity histogram of image at typical excitation

The histogram of the image is examined for locating peaks and valleys. A parameter *"θ"* called threshold is chosen and applied to the image *a* [*m*, *n*] as follows:

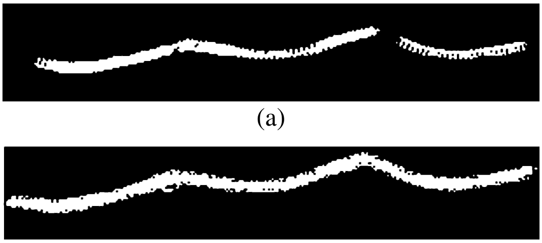
If $a\,[m, n] \geq \theta$ $a\,[m, n] =$ **Lamp Filament** = 1
Else $a\,[m, n] =$ **Background** = 0

The threshold is chosen from the brightness histogram of the region that is expected to segment [9].

Following observations are made after studying the brightness histogram:

The histogram is multimodal. The largest peak in the histogram depicts the background and smaller ones represent filament and reflection from shell. The peaks at higher brightness representing the filament are the objects of interest.

A threshold is selected iteratively by examining local neighbourhood histogram. This threshold value is used to filter out unwanted data. The filament images obtained after thresholding are shown in Fig. 21.9. Images (a) and (b) are the filament images at 100 and 120 V, respectively, obtained after thresholding.



(a)

(b)

Fig. 21.9. Filament images after thresholding

Intensity difference is observed in the above filament images. Filament temperature changes with change in intensity. The colour temperature of the filament can be found out with reference to black body radiations [9]. According to Stefan–Boltzmann's law, the energy emitted by a black body per unit area and unit time is proportional to the power "four" to the absolute temperature of the body. The "grey" body is represented by the filament of an incandescent lamp whose energy emission is investigated as a function of the temperature. The energy emitted per unit area and unit time at temperature T and wavelength λ within the interval dλ is designated by d$L(T, \lambda)$ / dλ. Planck's formula states,

$$\frac{\mathrm{d}L(T,\lambda)}{\mathrm{d}\lambda}=\frac{2c^{2}}{\exp(^{hc}/_{kT\lambda})}\frac{h\lambda^{-5}}{-1} \qquad (21.1)$$

where c= velocity of light (3×10^{8} m/s),

h= Planck’s constant (6.62×10^{-34}),
k=Boltzmann’s constant (1.381×10^{-23}).

Integration of this gives the relation of intensity and temperature, i.e. Stefan–Boltzmann’s law.

$$L(T) = \frac{2\Pi 5\, k^{4^{4}} T^{4}}{15\, c^{2} h^{3}} = \sigma T^{4} \tag{21.2}$$

where $\sigma = 5.67 \times 10^{-8}$.

On the basis of these laws, temperature can be related with the intensity of the source. Temperature–intensity co-relation is shown in Fig. 21.10. If these temperature values are calibrated using a standard, filament temperature can be predicted [9].

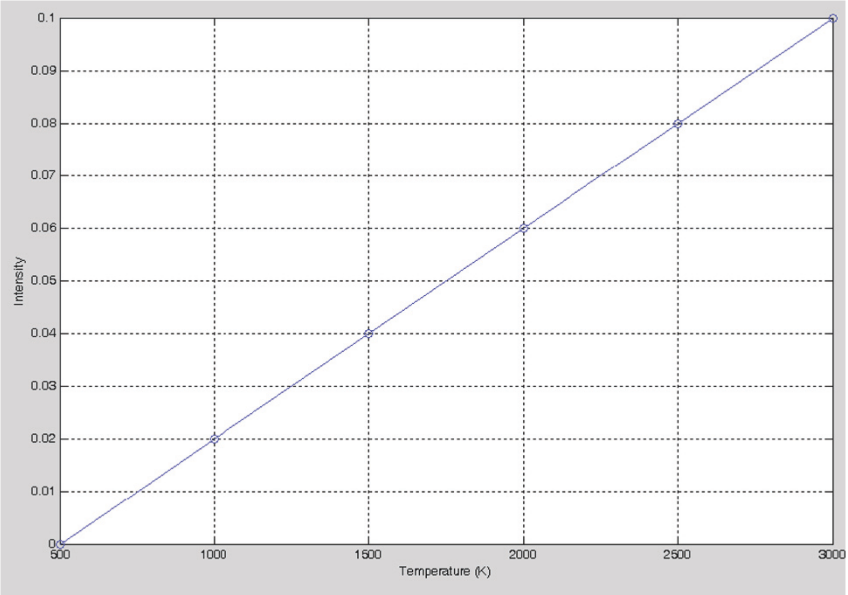


Fig. 21.10. Intensity–temperature co-relation

21.5 Results and conclusion

The optical characteristics of the camera used for image capturing are obtained. From these results, it is observed that the digital camera used to capture source images is having enough bandwidth to cover visible range (400–700 nm).

Filament images of a typical 40 W incandescent lamp are captured using digital camera of known resolution. Desired image segment is selected for processing using thresholding. Intensity change can be correlated with temperature. Using a standard, temperature–intensity relation is established. Using this relationship for a known intensity value, filament temperature can be predicted.

References

1. Wood M (2001) Color temperature of metal halide sources.
2. Rananath R, Snyder W, Yoo Y, Drew M (2005) Color image processing. IEEE Signal Magazine, pp 16–20
3. Stoerring M (1999) Light spectra and colour temperature.
4. http://www.wartsila-nsd.com
5. http://www.Colorphenomena.com
6. Gonzalez RC, Woods RE (2003) Digital image processing. Pearson Educatio, Singapore
7. Jain AK (2001) Fundamentals of digital image processing. Prentice Hall Private Limited, New Delhi
8. Horace, Diggang S (2002) An active surface paradigm for adaptive image thresholding.
9. Tilton JC Hierarchical image segmentation. NASA's Goddard space flight center, Greenbett, Maryland
10. Dooley JW (2002) Black body radiation and characteristics of incandescent lamp.

Chapter 22

Discrete decentralized observation of large-scale interconnected systems

M. Zazi[1], N. Elalami [2]

[1]Department Electrical, ENSET Rabat, BP 6217 Rabat Institut, Morocco, malkazazie@yahoo.fr
[2]Department Electrical of Engineering, Mohammadia School of Engineering Morocco, elalami@emi.ac.ma

Abstract. A new approach for the design of discrete decentralized observation schemes for large-scale interconnected systems is considered. The design is based on stability result that employs the notion of block diagonal dominance in matrices and the reasonable bound estimates for the discrete Lyapunov matrix equation. The major contribution of this chapter is the demonstration of how the observer's gains can be tailored systematically to the existing interconnection pattern within the overall system. Although the present results are developed in the context of decentralization observation, they can be extended to the design of decentralized stabilization and to the design of decentralized model reference adaptive identification schemes. Simulation results on a numerical example are given to verify the proposed design.

Keywords. Large-scale system, Diagonal dominance, Continuous Lyapunov equation, Discrete Lyapunov equation, Decentralized control

22.1 Introduction

Large-scale interconnected systems can be found in such diverse fields as electrical power systems, space structures, manufacturing process,

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_22,

transportation, and communication. An important motivation for the design of decentralized schemes is that the information exchange between subsystems of a large-scale system is not needed; thus, the individual subsystem's controllers are simple and use only locally available information.

Decentralized control of large-scale systems has received considerable interest in the systems and has benefit from numerous studies. Early work in the area can be found in [1–3]. More recently, Sundareshan and Elbanna [4,5] present a systematic constructive procedure based on a stability result that employs the notion of block -diagonal dominances in matrices. But the implementation of these controllers is very complicated and the obtained gains are very high.

To reduce the controller's gains, Elmarjany and Elalami [6] studied the decentralized stabilization via eigenvalues assignment and developed the sufficient condition under which exponential stabilization with prescribed convergence rate is achieved. The obtained gains are smaller than those found in other designs. Zazi and Elalami [7] extended the previous work to discrete-time systems and developed a new and simple approach for the design of discrete decentralized controllers of large-scale interconnected linear systems.

In many practical situations, complete state measurement is not available at each individual subsystem for decentralized control; consequently, one has to consider decentralized feedback control based on measurements only or design decentralized observers to estimate the state of individual subsystem that can be used for estimated state feedback control. There has been a strong research effort in literature toward the problem of designing observers. Some applications of these designs have been made to the observation problems arising in such diverse areas as spacecraft control and control of industrial manipulators.

An approach to decentralized observation that has yielded useful results [5] is to first construct a set of local observes for the independent subsystems and then to incorporate compensatory signals in order to account for the presence of interconnections among the subsystems. The objective of this chapter is to extend this work to discrete-time systems and develop a new and simple algorithm for the design of discrete decentralized observers of large-scale interconnected linear systems.

22.2 Problem formulation

Consider a large-scale discrete system s described as an interconnection of N subsystems, $s_1,s_2,\ldots,s_N$, by

$$x_i(k+1) = A_i x_i(k) + B_i u_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{N} H_{ij} x_j(k) \tag{22.1}$$

$$y_i(k) = C_i x_i(k) \qquad i = 1,\cdots,N$$

$x_i \in \Re^{ni}$ is the state of subsystem s_i, $u_i \in \Re^{mi}$ is its input vectors and $y_i \in \Re^{pi}$ is its output vectors.

$\sum_{\substack{j=1 \\ j\neq i}}^{N} H_{ij}$ is the term due to interconnection of the order subsystems. A_i, B_i and C_i are matrices of appropriate dimensions.

It is assumed that all pairs (A_i, B_i) are controllable and (A_i, C_i) are completely observable for all i,j=1, 2, …, N.

Let us also consider the observation scheme

$$\hat{x}_i(k+1) = (A_i - l_i C_i)\hat{x}_i(k) + B_i u_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{N} \Gamma_{ij} \hat{x}_j(k) \tag{22.2}$$

where $l_{ij} \in R^{n_i \times n_j}$ are the observer gains suitably selected.

When $H_{ij} \neq 0$, the selection of $\Gamma_{ij} = H_{ij}$ results in the dynamics of the estimation error being governed by

$$e_i(k+1) = (A_i - l_i C_i)e_i(k) + \sum_{\substack{j=1 \\ j\neq i}}^{N} H_{ij} e_j(k) \tag{22.3}$$

The problem of interest then is the choice of l_i such that the overall error system

$$e(k+1) = (A+H-1C)\ e(k) \tag{22.4}$$

where $e=[e_1^T\ e_2^T\ \ldots e_N^T]$, A=diag $[A_1 A_2 \ldots A_N]$, l=diag $[l_1\ l_2 \ldots . l_N]$, and C=diag $[C_1\ C_2 \ldots . C_N]$, is asymptotically stable.

Recognition of this relation between the stability of the error system and the design of the observation scheme enables one to employ the available results from the stability and the stabilization of large-scale systems to the observer design problem.

We first show how to find a local controller in the continuous- and discrete-time case.

Consider a large-scale continuous-time system:

$$\dot{x}(t) = A_i x_i(t) + B_i u_i(t) + \sum_{\substack{j=1 \\ j\neq i}}^{N} H_{ij} x_j(t) \tag{22.5}$$

$$y_i = C_i x_i(t) \qquad i = 1, \cdots, N.$$

The objective is to find a decentralized linear constant feedback control $u_i(t) = -K_i x_i(t)$ to exponentially stabilize the system (22.5).

Applying the i controller to the plant (22.5) gives

$$\dot{x}_i(t) = F_i x_i(t) + \sum_{\substack{j=1 \\ j\neq i}}^{N} H_{ij} x_j(t) \qquad i = 1, \ldots, N \tag{22.6}$$

$$F_i = A_i - B_i K_i.$$

Let $\lambda_{\min}(X)$ and $\lambda_{\max}(X) = \lambda_1(X)$, respectively, denote the minimum and the maximum of reel matrix X, the notations $\lambda(X)$ and $\sigma(X)$ denote the eigenvalue and singular value of the matrix X. Also for any $P \in \Re^{n\times m}$ $\|P\| = \sqrt{\lambda_1(P^T P)} = \sigma_1(P)$, let μ (.) denote the matrix measure induced by some vector or matrix norm and defined by the formula

$$\mu(A) = \lim_{\theta \to 0^+} \frac{\|I + \theta A\| - 1}{\theta}. \tag{22.7}$$

The matrix measure induced by the 2-norm is denoted by μ_2(A), and $\mu_2(A) = \frac{1}{2}\lambda_1(A + A^T)$. Moreover, the matrix measure μ_M (A) is given by

$$\mu_M(A) = \frac{1}{2}\lambda_1(MAM^{-1} + A^T) = \mu_2(M^{1/2} A M^{-1/2}). \tag{22.8}$$

Definition *1* Let $A \in \Re^{nxn}$ be partitioned in the form

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & & & \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \tag{22.9}$$

where $A_{ii} \in \Re^{ni\times ni}$ and $A_{ij} \in \Re^{ni\times nj}$, $i, j = 1, 2, \ldots, n$. If A_{ii} are non-singular and

$$\left\|A_{ii}^{-1}\right\|^{-1} \geq \sum_{\substack{j=1 \\ j\neq i}}^{n} \left\|A_{ij}\right\|, \tag{22.10}$$

then A is said to be block diagonal dominant relative to the partitioning in (22.9). If strict inequality holds in (22.10), then A is strictly block diagonal dominant.

Lemma 1 Let the matrix A partitioned as in (22.9) satisfy the conditions (i) $A=A^T$, (ii) A_{ii}=1, 2,….,n are positive definite, and (iii) A is strictly block diagonal dominant. Then, all the eigenvalues of A are real positive.

Lemma 2: Let M be a positive definite matrix satisfying μ_M (F_i) <0. Let spec (F_i) ⊂ LHP, ∀ i=1, 2,….,N and let P_i be the symmetric matrix solution of the Lyapunov equation

$$F_i^T P_i + P_i F_i + Q_i = 0. \tag{22.11}$$

For an arbitrarily selected symmetric matrix $Q_i \in \Re^{ni\times ni}$

$$\lambda_1(P_i) \leq \frac{\lambda_1(M)\lambda_1(M^{-1}Q_i)}{-2\mu_M(F_i)}. \tag{22.12}$$

In particular, if μ_2 $(A) < 0$, then we have

$$\lambda_1(P_i) \leq \frac{\lambda_1(Q_i)}{-2\mu_2(F_i)}. \tag{22.13}$$

Theorem 1 Let spec (F_i) ⊂LHP, i=1,2,…,N, and let P_i be the symmetric matrix solution of the Lyapunov equation (22.11). For an arbitrarily selected symmetric matrix $Q_i \in \Re^{ni\times ni}$, (22.5) is asymptotically stable if

$$\lambda_{\min}(Q_i) \succ \alpha_i \sum_{\substack{j=1 \\ j\neq i}}^{N} \left\|H_{ij}\right\| + \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_j \left\|H_{ji}\right\| \tag{22.14}$$

$$\alpha_i = \frac{\lambda_1(M)\lambda_1(M^{-1}Q_i)}{-2\mu_M(F_i)}$$

Proof: See [3]

The objective of this chapter is to extend this work to discrete systems, for that we must initially seek the sufficient condition for the existence of discrete decentralized controllers.

Let us consider a discrete large-scale system described as interconnections of N subsystems by

$$x_i(k+1) = F_i x_i(k) + \sum_{1}^{N} H_{ij} x_j(k) \tag{22.15}$$

where F_i is an asymptotically stable matrix.

Theorem 2: Let F_i be an asymptotically stable matrix, i=1, 2,…,N, and let P_i be the symmetric solution of the discrete algebraic Lyapunov matrix equation,

$$F_i^T P_i F_i - P_i + Q_i = 0 \tag{22.16}$$

For an arbitrarily selected symmetric matrix Q_i, then (22.15) is asymptotically stable if

$$\lambda_{\min}(Q_i) \succ \lambda_{\max}(P_i)\left\|F_i\right\| \sum_{\substack{j=1\\ j\neq i}}^{N} \left\|H_{ij}\right\| + \sum_{\substack{j=1\\ j\neq i}}^{N} \lambda_{\max}(P_j)\left\|H_{ij}\right\|\left\|F_j\right\| + \tag{22.17}$$

$$\sum_{\substack{k=1\\ k\neq i}}^{N}(\lambda_{\max}(P_k)\left\|H_{ki}\right\| (\sum_{\substack{j=1\\ j\neq i}}^{N} \left\|H_{kj}\right\|)) + \sum_{\substack{j=1\\ j\neq i}}^{N}\lambda_{\max}(P_j)\left\|H_{ji}\right\|^2$$

Proof Selecting $V(x(k))= x^T(k)\ P\ x(k)$, P=diag(P_1, P_2,…, P_N), as a discrete Lyapunov function and evaluating its variation along the trajectories of (22.15)

$$\begin{aligned} \Delta V(x) &= V(x(k+1)) - V(x(k)) \\ &= x^T(k)\ (F^T P H + H^T P F + H^T P H - Q\ )\ x(k) \\ &= -x^T(k)\ W\ x(k) \end{aligned} \tag{22.18}$$

where Q=diag(Q_1,Q_2,…,Q_N), F=diag(F_1,F_2,…,F_N), H=[H_{ij}]i,j=1,…,N (22.19)

$W=Q\text{-}F^T\ P\ H\text{-}H^T\ P^{\cdot\cdot}F\text{-}\ H^T\ P\ H$ satisfies $W=W^T$

The diagonal element

$$W_{ii} = Q_i - \sum_{j=1(j\neq i)}^{N} H_{ji}^T P_j H_{ji},\ i=1,2,\ldots,N.\ \ W_{ii}^T = W_{ii}$$

$$W_{ij} = -F_i^T P_i \, H_{ij} - H_{ji}^T P_j \, F_j \, - \sum_{\substack{k=1 \\ k \neq i}}^{N} H_{ki}^T P_k \, H_{kj}$$

It is simple to observer that for $i \neq j$

$$\left\| W_{ij} \right\| \leq \lambda_{\max}(P_i) \left\| F_i \right\| \left\| H_{ij} \right\| + \lambda_{\max}(P_j) \left\| H_{ji} \right\| \left\| F_j \right\| + \sum_{\substack{k=1 \\ k \neq i}}^{N} \lambda_{\max}(P_k) \left\| H_{ki} \right\| \left\| H_{kj} \right\|. \tag{22.20}$$

If condition (22.17) is checked, then $W_{ii} > 0$ and $\left\| W_{ii}^{-1} \right\|^{-1} \succ \sum_{\substack{j=1 \\ j \neq i}}^{N} \left\| W_{ij} \right\|$, from Lemma 1, W is positive definite. For more details see [7].

Many researchers have developed results of upper bounds for discrete Lyapunov matrix $F^T PF - P + Q = 0$. All the existing results are based on the assumptions of $\lambda_1 \; (F \; F^T) \prec 1$. This is obviously restrictive because the stability of F does not guarantee this assumption.

To cover the case that $\lambda_1 \; (F \; F^T)$ is not inside the unit circle, Dong-Gi Lee, et al. [8] use the similarity transformation and set

$$\hat{P} = T^T P \, T, \quad \hat{Q} = T^T Q \, T, \quad \hat{F} = T^{-1} F \, T. \tag{22.21}$$

Then the modified Lyapunov equation is obtained

$$(T^T F^T T^T)(T^T P \, T) \, (T^1 \, FT) - T^T P \, T + T^T Q \, T = 0. \tag{22.22}$$

Theorem 3 Let the positive-defined matrix P be the solution to (22.16), if $\sigma_1(\hat{F}) < 1$

$$\lambda_i(P) \leq \lambda_1(E) \left[\lambda_i \left[\frac{\lambda_1(E^{-1}Q)\hat{F}^T \hat{F}}{[1 - \sigma_1^2(\hat{F})]} + E^{-1} \, Q \right] \right] \tag{22.23}$$

$$1 \leq i \leq n, \quad E = T^{-T} T^{-1}$$

22.3 An algorithm for discrete decentralized observation

As discussed earlier, the error system is described by

$$e_i(k+1)=(A_i-l_iC_i)e_i(k)+\sum_{\substack{j=1\\ j\neq i}}^{N}H_{ij}e_j(k) \tag{22.24}$$

$$e_i(k)=x_i(k)-\hat{x}_i(k),\quad F_i{=}A_i-l_i\,C_i.$$

Theorem 4 Let F_i be an asymptotically stable matrix and let P_i be the solution of the discrete Lyapunov equation. For an arbitrarily selected Q_i, where Q$_{\text{i}}$ is a positive definite matrix, (22.24) is asymptotically stable if

$$\lambda_{\min}(Q_i)\succ\ \alpha_i\|F_i\|\sum_{j=1(j\neq i)}^{N}\|H_{ij}\|+\sum_{j=1(j\neq i)}^{N}\alpha_j\|H_{ij}\|\|F_j\| \tag{22.25}$$

$$+\sum_{\substack{k=1\\ k\neq i}}^{N}(\alpha_k\|H_{ki}\|\ (\sum_{\substack{j=1\\ j\neq i)}}^{N}\|H_{kj}\|))+\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_j\|H_{ji}\|^{2}$$

$$\alpha_i=\lambda_1(E)\left[\lambda_1\left[\frac{\lambda_1(E^{-1}Q_i)\hat{F}_i^{\,T}\hat{F}_i}{[1-\sigma_1^2(\hat{F}_i)]}+E^{-1}Q_i\right]\right]$$

Proof From Theorem 2, the system (22.24) is asymptotically stable if the condition (22.17) is satisfied. From Theorem 3, the solution has an upper bound (22.23), through which we obtain the following inequality:

$$\lambda_{\max}(P_i)\|F_i\|\sum_{\substack{j=1\\ j\neq i}}^{N}\|H_{ij}\|+\sum_{\substack{j=1\\ j\neq i}}^{N}\lambda_{\max}(P_j)\|H_{ij}\|\|F_j\|\ + \tag{22.26}$$

$$\sum_{\substack{k=1\\ k\neq i}}^{N}(\lambda_{\max}(P_k)\|H_{ki}\|(\sum_{\substack{j=1\\ j\neq i}}^{N}\|H_{kj}\|))+\sum_{\substack{j=1\\ j\neq i}}^{N}\lambda_{\max}(P_j)\|H_{ji}\|$$

$$\prec\ \alpha_i\|F_i\|\sum_{\substack{j=1\\ (j\neq i}}^{N}\|H_{ij}\|+\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_j\|H_{ij}\|\|F_j\|\ +$$

$$\sum_{\substack{k=1\\ j\neq i}}^{N}(\alpha_k\|H_{ki}\|\ (\sum_{\substack{j=1\\ j\neq i}}^{N}\|H_{kj}\|))+\sum_{\substack{j=1\\ j\neq i}}^{N}\alpha_j\|H_{ji}\|^{2}$$

The system (22.24) is asymptotically stable if the condition

$$\lambda_{\min}(Q_i) \succ \alpha_i \sum_{j=1(j\neq i)}^{N} \|F_i\| \|H_{ij}\| + \sum_{j=1(j\neq i)}^{N} \alpha_j \|H_{ij}\| \|F_j\| + \sum_{\substack{k=1\\k\neq i}}^{N} (\alpha_k \|H_{ki}\| (\sum_{\substack{j=1\\j\neq i)}}^{N} \|H_{kj}\|)) + \sum_{\substack{j=1\\j\neq i}}^{N} \alpha_j \|H_{ji}\|^2 \tag{22.27}$$

is satisfied. This condition is most robust than the condition (22.17).

The implementation of this observer by using Theorem 4 is very complicated because the resolution of equation of algorithm imposes constraining conditions on the interconnection matrices and leads to restricted classes of the interconnected system. Moreover, the obtained gains are very high. In this chapter, we propose a new and simple algorithm for decentralized observation.

Let us consider the following transformation

$$X_i(k) = \gamma^{-k} x_i(k), \qquad U_i(k) = \gamma^{-k} u_i(k) \tag{22.28}$$

γ is a positive scalar ($\gamma \geq 1$)

$$\hat{X}_i(k+1) = \gamma^{-(k+1)} \hat{x}_i(k+1) = \gamma^{-1}(A_i - L_i C_i)\hat{x}_i(k) + \gamma^{-1} B_i U_i(k) + \sum_{\substack{j=1\\j\neq i}}^{N} \gamma^{-1} H_{ij} \hat{X}_j(k). \tag{22.29}$$

Then the estimation error of the modified system is governed by

$$\varepsilon_i(k+1) = \gamma^{-1}(A_i - L_i C_i)\varepsilon_i(k) + \sum_{\substack{j=1\\j\neq i}}^{N} \gamma^{-1} H_{ij} \varepsilon_j(k). \tag{22.30}$$

The goal is to select the observation gain L_i and the required scalar γ, such that the overall error system (22.30) is asymptotically stable.

From Theorem 4, the system (22.30) is asymptotically stable if

$$\lambda_{\min}(Q_i) \succ \gamma^{-1} \alpha_i \|F_i\| \sum_{j=1(j\neq i)}^{N} \|H_{ij}\| + \gamma^{-1} \sum_{j=1(j\neq i)}^{N} \alpha_j \|H_{ij}\| \|F_j\| + \gamma^{-2} \sum_{\substack{k=1\\k\neq i}}^{N} (\alpha_k \|H_{ki}\| (\sum_{\substack{j=1\\j\neq i)}}^{N} \|H_{kj}\|)) + \gamma^{-2} \sum_{\substack{j=1\\j\neq i}}^{N} \alpha_j \|H_{ji}\|^2 \tag{22.31}$$

We shall now give the procedure for the construction of the observer gains l_i.

Step 1: γ =1 selected L_i such that spec (A_i–L_i C_i) is inside the unit circle. This selection can be made by a standard pole placement design.

Step 2: Choose an arbitrary matrix positive Q_i.

Step 3: If condition (22.31) is checked, then calculate l_{i1} such that spec (A_i–l_{i1} C_i) = spec (γ^{-1} (A_i–L_{i1} C_i)).

Step 4: If condition (22.17) is checked, then l_i =l_{i1}.

If not, $\gamma = \gamma$ +1 and go to step 3.

If not, $\gamma = \gamma$ +1 and go to step 3.

7.4 Illustrative example

We consider the following example which was treated by Sundershan and Elbanna [5].

The discrete-time model is obtained from its continuous-time model by discretizing it using MATLAB c2d with the sampling period T=0.1.

$$x_1(k+1)=\begin{bmatrix}1.0637 & 0.2013 & 0.4264 & 0.8956\\ 0.3275 & 1.0293 & 0.0616 & 0.1340\\ 0.2587 & 0.3050 & 1.4371 & 0.7653\\ 0.0142 & 0.0174 & 0.1502 & 1.1474\end{bmatrix}x_1+$$

$$+B_1u_1(k)+\begin{bmatrix}0.0310 & 0.0426 & -0.0327\\ 0.1431 & 0.0286 & -0.0322\\ 0.0072 & -0.0200 & 0.0165\\ -0.0049 & -0.0049 & 0.0088\end{bmatrix}x_2(k)$$

$$y_1(k)=\begin{bmatrix}1 & 0 & 2 & 0\end{bmatrix}x_1(k)$$

$$x_2(k+1)=\begin{bmatrix}1.0074 & 0.1082 & 0.0264\\ 0.1083 & 1.1697 & 0.5633\\ -0.5681 & -0.0286 & 1.2731\end{bmatrix}x_2(k)$$

$$+B_2u_2(k)+\begin{bmatrix}0.0053 & 0.0038 & 0.0249 & -0.2327\\ 0.2557 & 0.0086 & 0.6204 & -0.0337\\ -0.0547 & 0.0109 & 0.3322 & -0.0769\end{bmatrix}x_2(k)$$

$$y_2(k) = \begin{bmatrix} 0 & 1.2 & 0 \end{bmatrix} x_2(k).$$

The resulting local observer gain matrices defined in (22.24) are

$$l_1 = \begin{bmatrix} 1.7089 & -19.2365 & 1.0409 & 8.1873 \end{bmatrix}^T$$

$$l_2 = \begin{bmatrix} -1.0386 & 2.4032 & 4.1746 \end{bmatrix}^T.$$

For comparison, observed gains determined in [5] are

$$l_1 = \begin{bmatrix} -674 & -4424.2 & 379 & 1594.4 \end{bmatrix}^T$$
$$l_2 = \begin{bmatrix} -252.4 & 69.208 & 435.05 \end{bmatrix}^T$$

which are considerably higher.

From the simulation results, we can notice that each estimation error is asymptotically stable (Figs. 22.1 and 22.2).

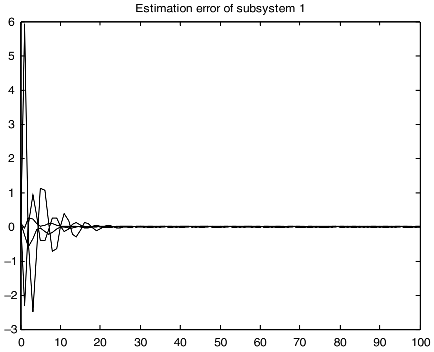


Fig. 22.1. Estimation error of the first subsystem

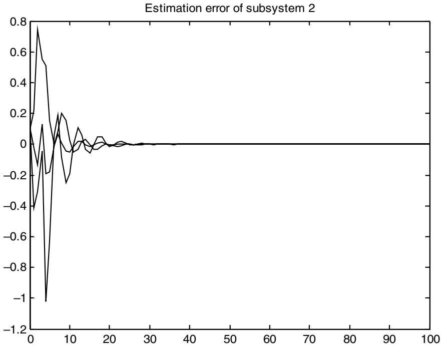


Fig. 22.2. Estimation error of the second subsystem

22.5 Conclusion

This chapter has presented a procedure for designing linear decentralized observer for discrete large-scale systems. Compared with existing results, our approach is more simple and easier to use and the gains obtained are smaller. The present results are developed in the context of decentralized observation; they have a wider application in that they can be extended to the design of decentralized model reference adaptive identification schemes. Future research is directed to the application of this approach to linear large-scale uncertain systems.

References

1. Sundershan M (1977) Exponential stabilization of large-scale systems; decentralized and multilevel schemes. IEEE Transaction System SMC-7:478–483
2. Siljac D (1978) Large Scale Dynamic Systems: Stability and Structure. Amsterdam, Netherlands
3. Soliman M, Darwich M (1978) Stabilization of large scale power systems via a multilevel technique. International Journal of System Science 9:1091–1111
4. Sundershan M, Elbanna R (1991) Constructive procedure for stabilization of large-scale system by informationnaly decentralized controllers. IEEE Transaction on Automatic Control 36(7):848–852
5. Sundershan M, Elbanna R (1990) Design of decentralized observation schemes for large-scale interconnected systems. Automatica 26(4):789–796
6. Elmarjany F, Elalami N (2004) Decentralized stabilization of large-scale systems with prescribed degree of exponential convergence. WSEAS Transaction Circuits and Systems 3(5):1178–1183
7. Zazi M, Elalami N (2006) Decentralized optimal stabilization of large-scale systems with prescribed degree of exponential convergence. WSEAS Transaction Circuits and Systems 5(8):1792–1796
8. Lee G, Hee Heo G, Myung Woom J (2003) New bounds using the solution of discrete lyapunov matrix equation. International Journal of Control and Systems 1(4)

Chapter 23

Improved iterative blind image deconvolution

Pankaj Kumar Sa[1], Ratnakar Dash[2], Banshidhar Majhi[3], Ganapati Panda[4]

[1] CSE, NIT Rourkela – 769008, India, pankajksa@gmail.com
[2] CSE, NIT Rourkela – 769008, India, ratnakar.dash@gmail.com
[3] CSE, NIT Rourkela – 769008, India, bmajhi@nitrkl.ac.in
[4] ECE, NIT Rourkela – 769008, India, gpanda@nitrkl.ac.in

Abstract. The simple technique of iterative blind deconvolution of two convolved functions has been improved in this chapter. The proposed improvement imposes some new constraints to the iterative algorithm making the result more accurate and more visually appealing. The proposed scheme has also been equipped with a stopping criterion ensuring the convergence of the algorithm which was missing in the original work. As a result, one need not look into the image produced after every iteration to decide the termination of the algorithm. Simulations indicate that the restored images of the proposed version of the scheme are more close to the true image.

Keywords. Image restoration, Blind image deconvolution, Point spread function, Motion blur

23.1 Introduction

Degradation of original image due to convolution is one of the frequently encountered problems in image processing. The convolution $g(x,y)$ of two functions, $f(x,y)$ and $h(x,y)$, can be written as

$$g(x,y) = f(x,y) * h(x,y) = \sum_{m,n} f(m,n)h(x-m,y-n) \tag{23.1}$$

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_23,

where $f(x,y)$ is the true image and $h(x,y)$ is the linear- shift-invariant blur known as point spread function (PSF) [1− 3].

If $F(u,v)$, $H(u,v)$, and $G(u,v)$ are the Fourier transforms of $f(x,y)$, $h(x,y)$, and $g(x,y)$, respectively, then Eq. (23.1) can otherwise be expressed as

$$G(u,v) = F(u,v)H(u,v) \tag{23.2}$$

Deconvolution is performed for image restoration in many applications such as astronomical imaging, medical imaging, and remote sensing. When one of the function $f(x,y)$ or $h(x,y)$ is known, the other function can be determined by inverse filtering or Wiener filtering. In classical linear image restoration problem, the PSF $h(x,y)$ is assumed to be known prior to the deconvolution process. To write in other words, when the blur function is known, the degradation process is inverted to get back the true image. However, in many practical situations the blur is often unknown, and little information is available about the original image [3−5].

Therefore, the image $f(x,y)$ must be identified directly from the convolved signal $g(x,y)$ using partial or no information about the blurring process and true image. Such an estimation problem, assuming the linear degradation model, is called blind deconvolution. Blind image restoration is the process of estimating both the true image and the blur from the degraded image characteristics using partial information about the imaging system. In this work, one such image restoration technique has been improved for better and definite result.

Rest of the chapter is organized as follows. The blur model is described in Section 23.2. The reported scheme for deconvolution is reviewed in Section 23.3. The proposed scheme is detailed in Section 23.4 followed by simulation results in Section 23.5. Finally, Section 23.6 provides the concluding remark.

23.2 Blur model

Suppose that a scene to be recorded undergoes a planer motion relative to the sensor. Assume the relative motion to be of uniform velocity v at an angle θ with the horizontal axis. If T is the duration of exposure, then the blur length is $L = vT$, and the motion blur PSF may be expressed as

$$h(x,y)=\begin{cases}1/L & \text{if } 0\le |x|\le L\cos\theta ; y=L\sin\theta \\ 0 & \text{otherwise}\end{cases} \tag{23.3}$$

If the motion is along horizontal direction, that is, $\theta = 0$, the above equation may then be expressed as

$$h(x,y)=\begin{cases}1/L & \text{if } 0\le |x|\le L ; y=0 \\ 0 & \text{otherwise}\end{cases} \tag{23.4}$$

23.3 Review of IBD

The iterative blind deconvolution (IBD) algorithm [5] iteratively estimates the original image as well as the PSF. IBD makes use of spatial domain as well as frequency domain constraints. In spatial domain, non-negativity constraint is used on both image as well as PSF. Non-negativity is used in spatial domain because image pixel intensity values are always positive. Similarly, PSF values are observed to be always positive. The Fourier domain constraint may be described as constraining the product of the Fourier spectra of $f(x,y)$ and $h(x,y)$ to be equal to the Fourier spectra of $g(x,y)$, in agreement with Eq. (23.2).

The basic deconvolution method consists of the following steps. First, a non-negative-valued initial estimate of the PSF $h(x,y)$ is input into the iterative scheme. The size of the PSF must be known before starting the algorithm. However, the PSF values can be random numbers. Now $h(x,y)$ is Fourier transformed to yield $H(u,v)$, which is then inverted to form an inverse filter and multiplied by $G(u,v)$ to form the first estimate of the original image spectrum $F(u,v)$. This estimated Fourier spectrum is inverse transformed to obtain $f(x,y)$. The image domain constraint of non-negativity is now imposed by putting zero to all pixels of the image $f(x,y)$ that have a negative value. A positive-constrained estimate $\hat{f}(x,y)$ is formed that is Fourier transformed to obtain the spectrum $\hat{F}(u,v)$. Now $G(u,v)$ is divided by $\hat{F}(u,v)$ to get the next spectrum estimate $H(u,v)$. A single iterative loop is completed by inverse Fourier transforming $H(u,v)$ to obtain $h(x,y)$ and by constraining this to be nonnegative, yielding the next PSF estimate $\hat{h}(x,y)$. The iterative loop is repeated until a satisfactory restored image is obtained.

The calculations involved here are quite straightforward and simple. However, the output is not definite, the algorithm may run into infinite loop without converging. In order to get the approximation of the true image the initial estimate of the PSF should not vary much with the original PSF, which in itself is a difficult task to realize in practical situation. The proposed scheme, described in the next section, is an improvement upon the IBD.

23.4 Proposed improvement

The drawbacks of the IBD are alleviated in the proposed scheme by imposing some more constraints. A convergence criterion has also been incorporated into the algorithm. The constraints used in the iterative deconvolution scheme are having substantial influences on the number of iterations required for the satisfactory restoration of the blurred image.

Support: Support is the smallest rectangle which encompasses the image area in the original image. So the pixel values outside the support are considered as the background color. After every iteration, the pixel values outside the support are made equal to the background color.

Non-negativity: Non-negativity constraint is applied on the image estimate and the PSF estimate. The negative values appearing in these estimates are replaced with zero. Also, the sum of the PSF values is made equal to 1 after every iteration.

$h_{\min}$: The spatial domain value of the PSF is always made more than the threshold value.

$$h(x,y)=\begin{cases} h_{\min} & \text{if } h(x,y)<h_{\min} \\ h(x,y) & \text{otherwise} \end{cases} \tag{23.5}$$

$F_{\max}$: The upper limit of the frequency domain values of the image are always set to $F_{\max}$.

$$F(u,v)=\begin{cases} F_{\max} & \text{if } F(u,v)>F_{\max} \\ F(u,v) & \text{otherwise} \end{cases} \tag{23.6}$$

$f_{\max}$: The upper limit of the spatial domain values of the image are fixed at $f_{\max}$.

$$f(x,y)=\begin{cases} f_{\max} & \text{if } f(x,y)>f_{\max} \\ f(x,y) & \text{otherwise} \end{cases} \tag{23.7}$$

These constraints are found to be indispensable for approximating the true image from the degraded observations.

23.4.1 Algorithm

I. Get the initial PSF $h_0(x,y)$ with random values.

II. Find $\hat{H}_k(u,v)$ by taking Fourier transform of $h_k(x,y)$.

III. Compute $F_k(u,v)$ from $G(u,v)$ and $\hat{H}_k(u,v)$ as $F_k(u,v)=G(u,v)/\hat{H}_k(u,v)$.

IV. Compute the inverse Fourier transform of $F_k(u,v)$ to obtain $f_k(x,y)$.

V. Impose the image constraint of non-negativity, support and $f_{\max}$ on $f_k(x,y)$ to obtain $\hat{f}_k(x,y)$.

VI. Obtain $\hat{F}_k(u,v)$ after Fourier transforming $\hat{f}_k(x,y)$.

VII. Compute $H_k(u,v)$ from $G(u,v)$ and $\hat{F}_k(u,v)$ as $H_k(u,v)=G(u,v)/\hat{F}_k(u,v)$.

VIII. Compute the inverse Fourier transform of $H_k(u,v)$ to obtain $h_k(x,y)$.

IX. Impose the $h_{\min}$ and PSF constraints on $h_k(x,y)$.

X. Compute power of the image for the kth iteration as

$$\mathrm{P}_k=\sum_{x=1}^{M}\sum_{y=1}^{N}\hat{f}_k(x,y)^2 \tag{23.8}$$

XI. Determine the standard deviation of power (σ) for the last s iterations. If the computed σ is less than a predetermined threshold value then stop the iteration, otherwise repeat from step II.

23.5 Simulation results

The proposed improved version of the iterative blind image deconvolution (IIBID) is simulated along with the reported iterative blind deconvolution (IBD) scheme. Number of iterations required for convergence and the peak signal to noise ratio (Eq. (23.9)) are the two performance metrics considered for the comparison. Simulations are carried out in MATLAB 7 in Intel Core 2 duo, 2.13 GHz machine.

$$\mathrm{PSNR} = 10\log_{10}\left(255^2/\mathrm{MSE}\right)\mathrm{d}B \tag{23.9}$$

$$\mathrm{MSE} = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}\left(f(x,y) - \hat{f}(x,y)\right) \tag{23.10}$$

where MN is the size of the image, and $f(x,y)$ and $\hat{f}(x,y)$ represent the pixel values at $(x,y)_{\mathrm{th}}$ location of original and restored image, respectively. Two different binary images are blurred and restored with both the IBD and the IIBID. The results are shown in Figs. 23.1 and 23.2.

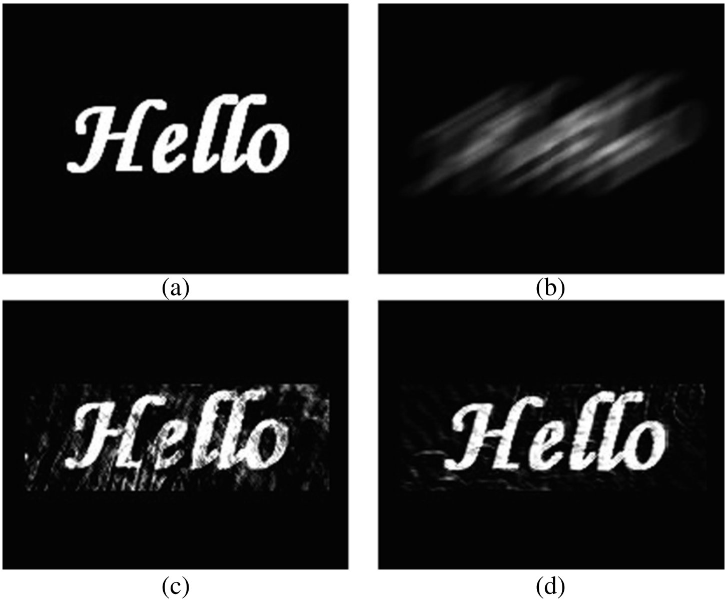
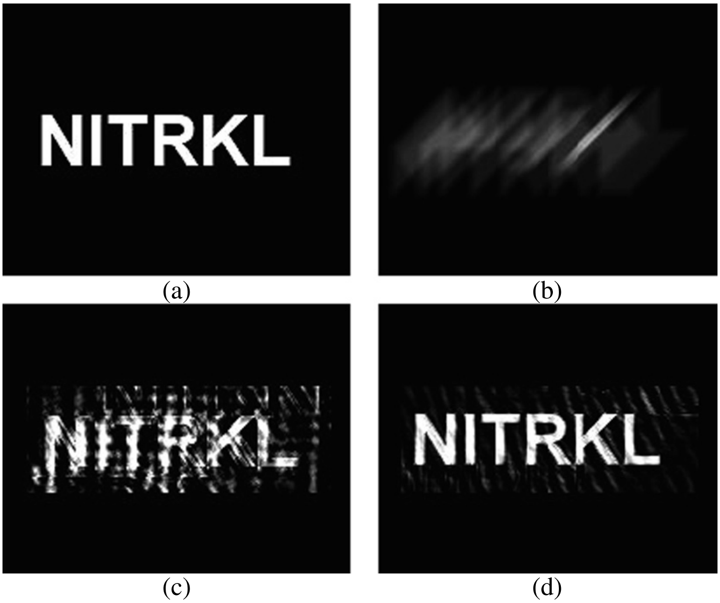

(a) (b) (c) (d)

Fig. 23.1. Restoration of *Hello* image. (**a**) True image, (**b**) motion blurred with $L = 60$, θ=30, (**c**) restored with IBD in 126 iterations (PSNR=17.13 dB), (**d**) restored with IIBID in 50 iterations (PSNR=20.57 dB)



(a) (b) (c) (d)

Fig. 23.2. Restoration of *NITRKL* image. (**a**) True image, (**b**) motion blurred with $L = 50$, $\theta = 45$, (**c**) restored with IBD in 638 iterations (PSNR=13.23 dB), (**d**) restored with IIBID in 206 iterations (PSNR=22.88 dB)

23.6 Conclusions

The improved iterative blind image deconvolution is an improvement of the classic iterative blind deconvolution scheme. The proposed scheme is incorporated with some more constraints to get a good restored image in reasonable amount of time. The scheme works very well for binary images, more suitable for astronomical imaging where images are obtained from a dark background. The improvement in the restored image quality is also substantial when compared with the earlier version.

References

1. Gonzalez RC, Woods RE (1992) Digital image processing. Addison Wesley, New York
2. Jain AK (1989) Fundamentals of digital image processing. Prentice-Hall of India, New Delhi

3. Kundur D, Hatzinakos D (1996) Blind image deconvolution. IEEE Signal Processing Magazine, pp 43–64
4. Sondhi MM (1972) Image restoration: The removal of spatially invariant degradation. IEEE 60:842–857
5. Ayers GR, Dainty JC (1988) Iterative blind deconvolution methods and its applications. Optics Letter 13:547–549

Chapter 24

Design of a linear quadratic Gaussian controller for a manufacturing process

M.K. Yurtseven,[1] B. Agaran[2]

[1] Systems Engineering Department, Yeditepe University, Kayısdagı, Istanbul, Turkey, kyurtseven@yeditepe.edu.tr
[2] Industrial Engineering Department, Dogus University, Zeamet sok, Istanbul, Turkey, bagaran@dogus.edu.tr

Abstract. In this study, the design procedure and the performance analysis of a linear quadratic gaussian controller (LQGC) are presented. The controller is expected to regulate the production−inventory process of a manufacturing system along a predetermined trajectory within the plant hierarchical control system; the hierarchical control structure is also described briefly in this chapter. The LQGC is based on a dynamic, discrete-time, and a stochastic model, aggregating products, and production processes. The simulation results show that the LQGC is well -suited for the intended purpose.

Keywords. Production planning and control, Linear quadratic Gaussian control, Hierarchical control, Systems engineering

24.1 Introduction

The concept of production planning and control and its dependence on the technology and the organizational forms were studied by many researchers in the past [1, 21, 7, 11, 13]. The most common finding of these studies is that the new technologies yield higher productivity, provided that they are implemented through appropriate organizational forms [1, 21]. This aspect is particularly emphasized in [21] where the contribution of new technolo-

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_24,

gies without sound management systems was found to be limited; the conclusions reported were based on an empirical study where four major factors of automation technology were evaluated for 15 machinery firms. Hence, a system engineering point of view of production planning and control appears to be more beneficial, compared to purely technological or purely managerial approaches. The present study is a typical example for the use of systems engineering techniques in handling production planning and control problems in a manufacturing process.

The application of control systems engineering techniques to manufacturing systems can be traced to [22,10,6] all of which fall within the body of classical control theory, which has its limitations when it comes to dealing with multi-input multi-output systems, time-varying systems, or non-linear systems. These were partially overcome with the application of modern control theory to the analysis and design of production−inventory systems, broadening the scope of applications [9,8,14]. The model developed in this study has been borrowed from control systems engineering area and was adopted earlier for production planning and control purposes [15,23]. Later, Yurtseven modified and extended the model to design a hybrid production planning and control system for a manufacturing process [24], and Yurtseven and Buchanan then proposed a similar model that could be used for assessing the effect of new technologies on production [25]. There has been a growing interest in the application of control system engineering techniques to the modeling and control of supply chain systems. The work reported by Perea et al. employs some ideas from process control to modeling and control of supply chains [20]. Lin et al. report a controller design study and its use on the reduction of bullwhip for a model supply chain [19]. The modeling approach is based on the Z-transform and the controller design is achieved in the frequency domain. Hoberg et al. apply linear control theory to study the effect of various policies on order and inventory variability which is considered to be the key drivers of supply chain performance [16]. Agaran, Buchanan, Yurtseven believe that the dominant dynamic characteristics of a complex system, such as a supply chain or a complex production–inventory system, can be modeled and controlled effectively with the powerful analytical tools of modern control theory, as opposed to the classical control theory [2]. They state the advantages of modern control theory over its classical counterpart: the latter is limited to the analysis of relatively simple systems that are linear, time-invariant, and small dimensioned (i.e., with small number of inputs and outputs). In modern control theory one can handle large-scale systems with several inputs and outputs without too much difficulty. In addition, the powerful techniques developed for linear and time-invariant systems can be extended to non-linear, time-varying, and stochastic systems effectively. In addition, it is possible to filter stationary or non-stationary noise present in signals through high-performance filters such as Kalman filters, design optimal control policies, and make use of adaptive techniques to update model parameters and control policies for more effective control.

The control-theoretic approach, like the other analytical tools, suffer from a major disadvantage; it is almost impossible to formulate complex issues such as organizational resistance to change, inter-functional or inter-organizational conflicts, team-oriented performance measures, customer relationship management etc., adequately. Min and Zhou suggest that the analytical tools alone are not sufficient to represent the realities of complex systems [17]. According to them, the traditional mathematical programming techniques can be used to model inter-functional integration, but realistic representations of such systems can be found through IT-based models that make use of model-based decision support systems (DSS). Such DSS have the potential of representing all the analytical and non-analytical aspects of complex systems in a more realistic manner. Hence, the work reported here is seen as part of an ongoing research where the overall objective is to develop a DSS for managing the manufacturing system under study. In other words, the model and the associated controller developed in this study will be a part of a DSS; it will be integrated with some other analytical/non-analytical tools within the DSS to cope with the ill-structured, strategic, and behavioral issues involved in the system.

The work reported in this chapter will be presented in the following order: The principles adopted in the design of the overall hybrid production planning and control system will be summarized in the next section. Descriptions of the plant model, the LQG controller structure, and the controller design procedure will follow this.

24.2 Principles of design

The reader will find here only a summary of the ideas considered during the design of the hybrid production and planning control system; details can be found in [24]. The hybrid control structure proposed is shown in Fig. 24.1. This structure is based on a concept developed by Kohn et. al. [18]. The plant or process under consideration is a workshop. The production strategies developed by the top management are translated into a set of production targets by Translator I and then fed into the production planning unit. Typically, weekly production plans are prepared within this unit. Note that the function of Translator I is to formulate production strategies set by the top management, which may be a mixture of quantitative and qualitative statements, into quantitative production targets in a specific format. Translator II translates these production plans into a specific form acceptable by the Scheduler. In turn, the Scheduler has the task of producing specific, typically daily schedules. Translator III transforms these schedules into specific control settings or production trajectories that are used by the controller. The hybrid nature of the control structure provides the "glue" between the event-based systems and continuum systems in the control hierarchy. The design of the control hierarchy, with its coherent control objectives and coordination schemes, requires a formal design procedure. The models that are used for production planning, scheduling, and controlling activities will have to be different; they normally have an increasing size, complexity, and level of detail as one goes down the hierarchy. Similarly, the time horizon considered by these models will need to decrease with decreasing level of hierarchy. Some discussion related to this topic can be found in [18, 5].

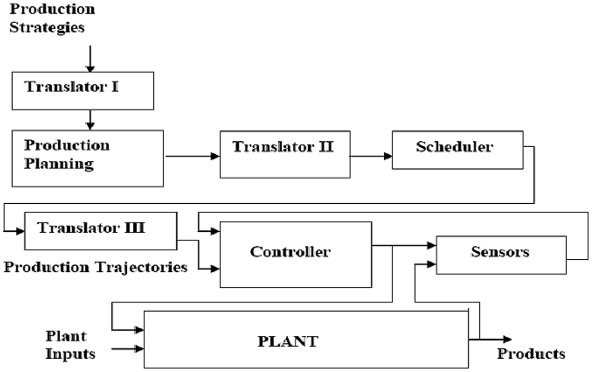


Fig. 24.1. The proposed hybrid production planning and control structure

24.3 The controller structure

The objective of the controller is to ensure that the production schedules prepared by the Scheduler are implemented properly. The controller may also be used as a pre-planning tool, providing the opportunity to systems engineers to test the possible contributions of the newer technologies and/or organizational forms into production [24]. The controller design is based on a dynamic model, providing the opportunity to investigate the variations in production under different control policies, at different production stages, as time progresses. The model is a discrete-time type, hence well suited to the discrete nature of the manufacturing process. Furthermore, its stochastic nature allows the systems engineer to incorporate the uncertainties involved in the process, providing some flexibility in the modeling of such complex phenomena. In order to keep the model at a reasonable size, products and production processes are aggregated. The aggregate aspect of the model allows the systems engineer to suppress the details and bring out the dominant characteristics of the production process, providing a systems perspective.

A block diagram of the LQG controller is shown in Fig. 24.2. The input to the controller is the vector of production trajectories. The controller generates the optimum control vector with components of $u_{11}(k)$, $u_{12}(k)$, $u_{21}(k)$, and $u_{22}(k)$, in period k. The former two represent increased or decreased number of machines at stages 1 and 2, respectively, and the latter two are the amount of overtime or under-time work exercised, at stages 1 and 2, respectively. The plant output $\boldsymbol{y}(k)$ is the available measurements. Due to the difficulties and cost involved in the measurement process, it is assumed that only $x_{21}(k)$ and $x_{22}(k)$ can be measured, which are the inventory levels at stages 1 and 2, respectively, in period k. A Kalman estimator is employed to estimate $\boldsymbol{x}_{1e}$ and $\boldsymbol{x}_{2e}$, which are the best estimates of $\boldsymbol{x}_1$ and $\boldsymbol{x}_{2,}$ respectively. The vector $\boldsymbol{r}_1(k)$ is a stochastic variable representing the unpredictable variations in the number of disabled or repaired machines during the period k. Similarly, the vector $\boldsymbol{r}_2(k)$ is another stochastic variable, representing the unpredictable variations in demand to the products in period k.

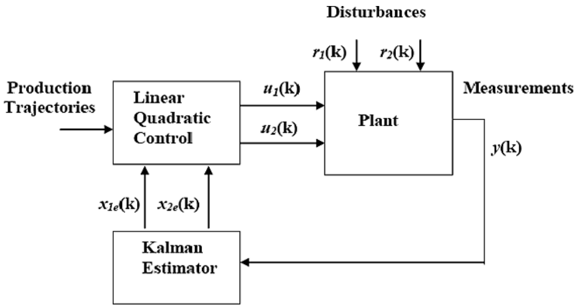


Fig. 24.2. A block diagram of the LQG controller

The objective of the controller is to regulate the plant around the nominal operating conditions. The linear model equations that represent small deviations from the nominal operating conditions are given by [23].

The transition of the number of machines in two successive periods is given by

$$\boldsymbol{x_1}(\mathrm{k}+1) = \boldsymbol{x_1}(\mathrm{k}) + \boldsymbol{u_1}(\mathrm{k}) + \boldsymbol{r_1}(\mathrm{k}),\quad \mathrm{k} = 0, 1, 2,\ldots, \mathrm{N}-1 \tag{24.1}$$

where the variables are as defined above.
The corresponding vectors are defined in the forms of
$\boldsymbol{x_1}(\mathrm{k}) = (\boldsymbol{x_{11}}(\mathrm{k}),\ldots, \boldsymbol{x_{1j}}(\mathrm{k}))'$
$\boldsymbol{u_1}(\mathrm{k}) = (\boldsymbol{u_{11}}(\mathrm{k}),\ldots, \boldsymbol{u_{1j}}(\mathrm{k}))'$
$\boldsymbol{r_1}(\mathrm{k}) = (\boldsymbol{r_{11}}(\mathrm{k}),\ldots, \boldsymbol{r_{1j}}(\mathrm{k}))'$
where $\boldsymbol{x_{1j}}(\mathrm{k})$, $\boldsymbol{u_{1j}}(\mathrm{k})$, and $\boldsymbol{r_{1j}}(\mathrm{k})$ represent these quantities at the jth production stage in period k = 0, 1, 2,…, N−1. Note that the symbol ′ indicates a matrix transposition operation.

The inventory level at the beginning of period k is given by

$$\boldsymbol{x_2}(\mathrm{k}+1) = \boldsymbol{x_2}(\mathrm{k}) + \boldsymbol{p}(\mathrm{k}) - \boldsymbol{r_2}(\mathrm{k}),\quad \mathrm{k} = 0, 1, 2,\ldots, \mathrm{N}-1 \tag{24.2}$$

where the variables are as defined earlier. The vectors have the forms of:
$\boldsymbol{x_2}(\mathrm{k}) = (\boldsymbol{x_{21}}(\mathrm{k}),\ldots, \boldsymbol{x_{2i}}(\mathrm{k}))'$
$\boldsymbol{p}(\mathrm{k}) = (\boldsymbol{p_1}(\mathrm{k}), \ldots, \boldsymbol{p_i}(\mathrm{k}))'$
$\boldsymbol{r_2}(\mathrm{k}) = (\boldsymbol{r_{21}}(\mathrm{k}),\ldots, \boldsymbol{r_{2i}}(\mathrm{k}))'$
$\boldsymbol{x_{2i}}(\mathrm{k})$, $\boldsymbol{p_i}(\mathrm{k})$, and $\boldsymbol{r_{2i}}(\mathrm{k})$ represent these quantities at the ith production stage in period k = 0, 1, 2,…, N−1.

The relation between the production time and the amount of products can be described by a linear approximation as

$$\mathbf{t}\,\boldsymbol{p}(\mathrm{k}) = \boldsymbol{b_p}(\mathrm{k}) \tag{24.3}$$

where $\boldsymbol{b_p}(\mathrm{k})$ is the production time in period k with $\boldsymbol{b_p}(\mathrm{k}) = (\boldsymbol{b_{p1}}(\mathrm{k}),\ldots, \boldsymbol{b_{pj}}(\mathrm{k}))'$, $\boldsymbol{b_{pj}}(\mathrm{k})$ representing this quantity at production stage j; **t** is the machining matrix with **t** = $(\mathrm{t_{ij}})$ with a dimension of jxi; $\mathrm{t_{ji}}$ represents the time required to produce one unit of product i (=1,2,…,I) at stage j (=1, 2,…, J).

The amount of products can then be written as

$$\boldsymbol{p}(\mathrm{k}) = (\mathbf{t}^{+})\,\boldsymbol{b_p}(\mathrm{k}) \tag{24.4}$$

where $(\mathbf{t}^{+})$ is the pseudo inverse of **t**. **t** is a square matrix when the number of products is equal to the number of production stages, and its inversion is easy. However, in some cases this is not a square matrix and its inverse has to be calculated through a special algorithm [23].

The production time in period k is calculated as follows:

$$\boldsymbol{b_p}(\mathrm{k}) = \mathbf{r_p}\ \boldsymbol{x_1}(\mathrm{k}) + \boldsymbol{u_2}(\mathrm{k}),\quad \mathrm{k} = 0, 1, 2,\ldots, \mathrm{N}-1 \tag{24.5}$$

where $\boldsymbol{u_2}(\mathrm{k})$ is the vector of overtime or under-time with $\boldsymbol{u_2}(\mathrm{k}) = (\boldsymbol{u_{21}}(\mathrm{k}),\ldots, \boldsymbol{u_{2j}}(\mathrm{k}))'$, $\boldsymbol{u_{2j}}(\mathrm{k})$ representing this quantity at production stage j. $\mathbf{r_p}$ represents the regular working time matrix with $\mathbf{r_p}$ = diagonal ($\mathbf{r_{pj}}$), with a dimension of j×j, $\mathbf{r_{pj}}$ being the regular working time at stage j = 1, 2,…, J.

Substituting Eq. (24.4) and (24.5) into (24.2) yields

$$\boldsymbol{x_2}(\mathrm{k}+1) = \boldsymbol{x_2}(\mathrm{k}) + (\mathbf{t}^{+})\,\mathbf{r_p}\,\boldsymbol{x_1}(\mathrm{k}) + (\mathbf{t}^{+})\,\boldsymbol{u_2}(\mathrm{k}) - \boldsymbol{r_2}(\mathrm{k}) \qquad (24.6)$$

$$\mathrm{k} = 0, 1, 2,\ldots, \mathrm{N}-1$$

The vector-matrix form of Eq. (24.1) and (24.6) can be written as follows:

$$\boldsymbol{x}(\mathrm{k}+1) = \mathbf{a}\,\boldsymbol{x}(\mathrm{k}) + \mathbf{b}\,\boldsymbol{u}(\mathrm{k}) + \mathbf{c}\,\boldsymbol{r}(\mathrm{k}), \quad \mathrm{k} = 0, 1, 2,\ldots, \mathrm{N}-1 \qquad (9.7)$$

where

$x(k) = [x_1(k)\;\; x_2(k)]'$, $u(k) = [u_1(k)\;\; u_2(k)]'$, $r(k) = [r_1(k)\;\; r_2(k)]'$

$a = [\mathrm{I}\;\; 0;\; (t^{+})r_p\;\; \mathrm{I}]$, $b = [\mathrm{I}\;\; 0;\; 0\;\; (\mathrm{t}^{+})]$, $c = [\mathrm{I}\;\; 0;\; 0\;\; -\mathrm{I}]$

Note that the symbol′ denotes matrix transposition, as defined earlier, and the symbol; separates the rows of the corresponding matrix (as used in MATLAB)

24.3.1 The design procedure

The mathematics of the LQG control is well known; hence they will not be repeated here. Instead, the design approach adopted and the criteria used in the selection of the critical design parameters will be explained, followed by a discussion on how simulation experiments were performed, and the results obtained. The reader will find the full information related to LQG control design in [3, 12, 4]. The system vector-matrix equations and data that are used during the simulation studies for various purposes are given in the Appendix. All design and simulation studies were performed using MATLAB.

The solution to the stochastic optimal control problem at hand is found through the well-known separation theorem or certainty equivalence Principle. According to this theorem, first an optimum estimator estimates the states of the model, ignoring the optimum control problem, and optimum control is then computed treating the estimated states as deterministic quantities. A *two product–two stage* case is considered in this study, with the following data:

$$\mathbf{t} = \begin{bmatrix} 1 & 0.5 \\ 3 & 2 \end{bmatrix} \qquad \mathbf{r_p} = \begin{bmatrix} 4 & 0 \\ 0 & 10 \end{bmatrix}$$

Here, **t** represents the machining time matrix, and $\mathbf{r_p}$ represents the regular working time matrix. The reader should note that $\mathbf{r_p}$ is a diagonal matrix, whereas **t** is an off-diagonal matrix, as expected. The plant state-space and measurement equations are put into the following standard form to be able to perform the design:

$$\boldsymbol{x}\,(\mathrm{k}+1) = \mathbf{A}\,\boldsymbol{x}(\mathrm{k}) + \mathbf{B}\,\boldsymbol{u}(\mathrm{k}) + \mathbf{G}\,\boldsymbol{w}(\mathrm{k}) \tag{24.8}$$

$$\boldsymbol{y}(\mathrm{k}) = \mathbf{C}\,\boldsymbol{x}(\mathrm{k}) + \mathbf{D}\,\mathrm{u}(\mathrm{k}) + \mathbf{H}\,\mathrm{w}(\mathrm{k}) + \boldsymbol{v}(\mathrm{k}) \tag{24.9}$$

where $\boldsymbol{w}$ (k) and $\boldsymbol{v}$ (k) are the random processes associated with process noise and measurement noise, respectively. **A**, **B**, **G**, **C**, and **D** are the corresponding system matrices, as given in the Appendix. **C** was chosen so that only the third and fourth state variables are available for measurement. The reader should note that the uncontrolled plant is unstable.

First, the (deterministic) LQ controller was designed, ignoring the noise processes. This was achieved through the use of MATLAB's *dlqr* command. The optimal control law is then computed to minimize the loss function $J_{c,}$ where $J_c = (\, \boldsymbol{x'}\, \mathbf{Q}_c\, \boldsymbol{x} + \boldsymbol{u'}\, \mathbf{R}_c\, \boldsymbol{u}\,)$. $\mathbf{Q}_c$ and $\mathbf{R}_c$ represent the weighting matrices for the state and control vectors, respectively. Several combinations of $\mathbf{Q}_c$ and $\mathbf{R}_c$ were simulated in order to tune the controller's performance. First, the *steady-state* optimal control law was calculated through the command *dlqr.* The Kalman estimator was designed through MATLAB's KALMAN command. The execution of this command requires the formulation of a quadratic loss function, similar to the one given above. This loss function contains $\mathbf{Q}_n$ and $\mathbf{R}_n$, which are the process noise and measurement noise covariance matrices, respectively. Once again, their values were chosen after some tedious tuning studies. A *steady-state* Kalman estimator was designed, as opposed to a time-varying one, since it satisfies the requirements of the regulator under consideration. The reader should also note that the stochastic variables $\boldsymbol{r_1}$(k), given in Eq. (24.1), and $\boldsymbol{r_2}$(k), given in Eq. (24.2), are included into the expression $\mathbf{G}\boldsymbol{w}$(k).

24.4 Conclusions

The design and performance analysis of a LQG controller for a complex manufacturing system were presented. The controller is intended to operate in a *hybrid* production planning and control structure where three translators serve as the "glue" between various subsystems of the production planning, scheduling, and control activities in the structure. In this chapter, a brief description of the proposed hybrid control structure was given, and the design procedure and performance analysis of the LQG controller was presented, fully. It was shown how products and production stages can be aggregated to construct a dynamic, discrete-time, and a stochastic model and how a LQG controller can be designed. Preliminary simulation studies conducted show that the resulting controller is able to regulate the plant under considerable noise or uncertainty reasonably successfully. Furthermore, more research needs to be conducted in the direction of designing the remaining components of the hybrid production planning and control system and test system's performance under realistic operating conditions.

References

1. Abernathy WJ, Townsend PL (1985) Technology, productivity and process change. In: Rhodes D, Wield E (eds) Implementing New Technologies, Basil Blackwell, Oxford
2. Agaran B, Buchanan WW, Yurtseven MK (2007) Regulating bullwhip effect in supply chains through modern control theory. Proceedings PICMET'07 Portland, USA
3. Anderson BDO, Moore JB (1990) Optimal Control Linear Quadratic Methods. Prentice Hall, New Jersey
4. Astrom KJ, Wittenmark B (1990) Computer-Controlled Systems: Theory and Practice. Prentice Hall, New Jersey
5. Bencze WJ, Franklin G (1995) A separation principle for hybrid control system design. IEEE Control System Magazine 15:80–85
6. Bishop AB (1975) Introduction to Discrete Linear Controls: Theory and Application. Academic Press, New York
7. Blumberg M, Gerwin D (1985) Coping with advanced manufacturing technology, In: Rhodes D, Wield E (eds) Implementing New Technologies, Basic Blackwell, Oxford
8. Christenson JL, Brogan WL (1971) Modeling and optimal control of a production process. International Journal of Systems Science 1:247–255
9. Connors CL, Teichroew D (1961) Optimal Control of dynamic operations research models. International Textbook Company, Pennsylvania
10. Elmagrahraby SEA(1966) The Design of Production Systems. Reinhold, New York
11. Ferraz JC, Rush H, Miles I (1992) Development, Technology and Flexibility: Brazil Faces the Industrial Divide, Routledge, London
12. Friedland B (1986) Control System Design: An Introduction to State-Space Method. McGraw Hill, New York
13. Gerwin D, Kolodny H(1992) Management of Advanced Manufacturing Technology. Wiley, New York
14. Hendricks CL, Koivo AJ (1971) An introduction to determine optimal policies for production-inventory systems. International Journal of Control, 14:341–351

15. Hitomi K, Nakamura M (1976) Optimal production planning for multiproduct-multistage production systems. International Journal of Production Research 14:199–213.
16. Hoberg, K, Bradley JR, Thonemann UW (2007)Analyzing the effect of the inventory policy on order and inventory variability with linear control theory European Journal of Operational Research 176:1620–1642
17. Hokey M, and Zhou G (2002) Supply chain modeling: past, present, and future. Computers & Industrial Engineering 43:231–249
18. Kohn W, James J, Nerode A, Agrawala A, Harbison K, (1995) A hybrid systems approach to computer-aided control engineering. IEEE Control Systems Magazine 15:14–25
19. Lin P-H, Wong S-H, Jang S-S, Shieh S-S, Chu J-Z (2004) Controller design and reduction of bullwhip for a model supply Chain system using z-transform analysis. Journal of Process Control 14:487– 499
20. Perea E, Grossmann I, Ydstie E, Tahmassebi T (2000) Dynamic modeling and classical control theory for supply chain management. Computers and Chemical Engineering 24:1143–1149
21. Rothwell S, Davidson D (1986) Manpower Matters: Technological Change, Company Personnel Policies and Skill Deployment, IFS Publications, Bedford
22. Simon HA (1952) On the application of servomechanism theory in the study of production control. Econometrica 2:247–268
23. Yurtseven MK, Bak T (1976).Time-optimal production control in a manufacturing system. International Journal of System Science 18:2175–2182
24. Yurtseven MK (1997) Design of a hybrid production planning and control system for a manufacturing process. Proceedings of the 2nd Asian Control Conference, pp 77–80
25. Yurtseven MK, Buchanan WW (1999) A model for assessing the effect of new technologies on production. PICMET'99, Proceedings, Vol-1, Book of Summaries, p 5

Chapter 25

3D reconstruction and isometric representation of engineering drawings

Muhammad Abuzar Fahiem[1,2], Anita Malik[1]

[1] University of Engineering and Technology, Lahore, Pakistan
[2] Lahore College for Women University, Lahore, Pakistan,
buzar@uet.edu.pk

Abstract. Engineering drawings are mostly represented in drawing exchange file (DXF) format for information interchange. DXF format is recognized by CAD tools only and results in a large file size, requiring heavy loading time. In this chapter, we have extended our previous work for the 3D reconstruction of engineering drawings in DXF and a new aspect is added to represent isometric views of these drawings in SVG format. The discussion is concluded on a comparison of SVG with DXF, proving the former suitable, specially for World Wide Web.

Keywords. Isometric views, SVG, DXF, 3D reconstruction, Engineering drawings

25.1 Introduction

Engineering objects are represented through engineering drawings from three standard orthographic views. These engineering drawings are used in the industry for manufacturing, machining, and production of engineering components. These drawings may be manual or computerized. Different computer-aided designing (CAD) tools have evolved in the past decades to develop computerized drawings. However, the computerization of older manual drawings is a vital research area addressed by different researchers.

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_25,

The main research focus is on the vectorization of these manual drawings in raster formats from camera perspectives or scanned projections. Converting manual drawings into a format supported by various CAD applications is also an area of interest to researches. The most widely used format supported by CAD tools for information interchange is drawing exchange file (DXF) format. DXF is basically a vector format. Representation of engineering documents on World Wide Web has recently gained popularity in the research circles. Scalable vector graphics (SVG) is an XML-based standard [1] of World Wide Web Consortium (W3C) evolved to represent different graphical objects across World Wide Web. Yet another active research area is to produce 3D models from these 2D orthographic projections. Different 3D reconstruction techniques have been evolved in the past decades for this purpose. In this chapter, we are dealing with 2D drawings and the output is a 3D drawing fully editable in DXF format. Moreover, our approach is capable of producing 3D isometric views in SVG format. A survey on different 3D reconstruction techniques is discussed in our previous work [2, 3]. Our approach for 3D reconstruction is discussed in our previous papers [4, 5], a summary of which is presented in Section 25.2.

Different algorithms have been proposed for raster to SVG conversion mostly depending on data-dependent triangulation (DDT), wavelet-based triangulation (WBT), and watershed decomposition (WD). DDT approach is discussed in [6], while WBT is used in [7]. WD technique for raster to SVG conversion is presented in [8, 9]. A good comparison of these techniques is discussed in [9, 10].

Another issue while dealing with the representation of engineering objects in SVG format is that engineering objects are nested into each other. It complicates the situation when arbitrary view points or page sizes are to be supported in a Web browser. Scaling is not guaranteed to maintain relative association of the objects loaded in pages of different sizes. A solution to this problem is constraint SVG [11–13].

While representing engineering documents over World Wide Web, different flavors of browsers [14] are to be handled. Bandwidth of the medium imposes serious limitations on the size of the document to be transmitted between different sites. Vector format SVG with its small file size plays a very important role in this regard; however, a suitable compression technique [15] can also be used.

Section 25.2 demonstrates our approach while Section 25.3 is dedicated to a comprehensive comparison of DXF and SVG formats from different critical aspects. The discussion is summarized in Section 25.4 followed by future recommendations in Section 25.5.

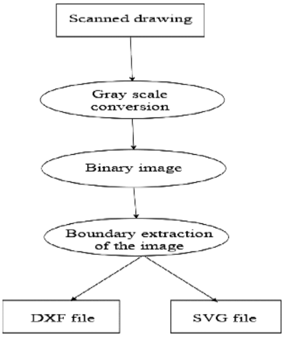


Fig. 25.1. Execution sequence of our approach

25.2 3D Reconstruction and isometric representations

Here, we use scanned manual drawings as input. These drawing images are binarized and gray scaled followed by a boundary trace. Our system is capable of generating output in both DXF and SVG formats. The flow of sequences is depicted in Fig. 25.1. DXF is generated by using our previous approach [5]. For isometric representation in SVG, the pseudocode is as follows:

```
Input: Scanned drawing in bmp
           Top view
           Front view
           Side view

1.  Grayscale conversion
2.  Binarization
3.  Boundary Extraction
4.  Rotate Top view about x-axis at an angle of $90^0$ clockwise.
5.  Rotate Top view about z-axis at an angle of $30^0$ clockwise.
6.  Rotate Side view about y-axis at an angle of $90^0$ clockwise.
7.  Rotate Side view about z-axis at an angle of $30^0$ clockwise.
8.  Rotate Front view about z-axis at an angle of $30^0$ clockwise.
9.  Translate Boundaries to align with the hypothetical cuboid as presented in [5].
10. Convert these boundaries in SVG entities.

Output: SVG file
```

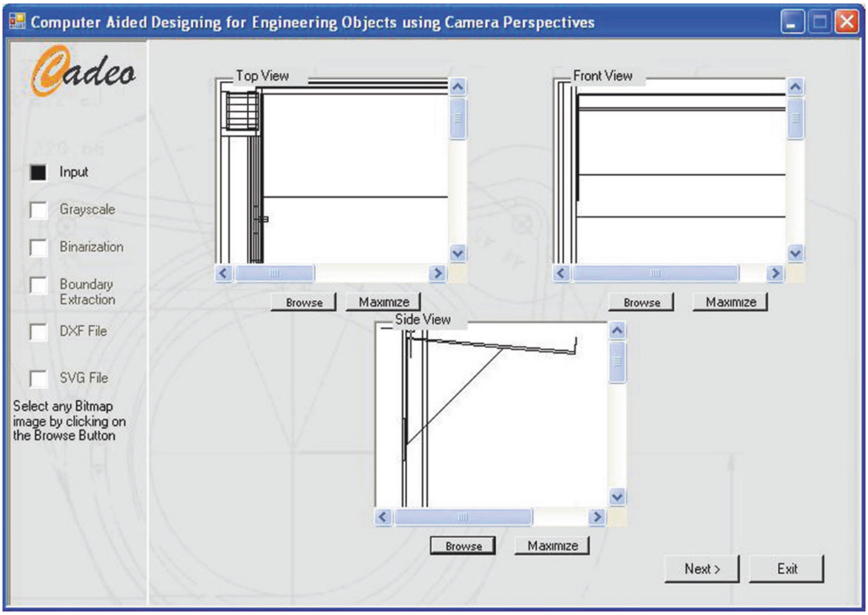
Here, first of all scanned projections of engineering drawings, namely top, side, and front, are taken as input. Then the operations from 1 through

10 are performed. While binarizing the image, threshold value can be specified by the user. So the user has a great control over choosing the details of the drawings. The output is generated as an SVG file viewable in any Web browser with suitable plug-ins.

The interface and the interaction of our application are shown in Fig. 25.2. Figure 25.3 is an AutoCAD 2005 view of a sample 3D drawing in DXF format generated by our algorithm, while Fig. 25.4 is the SVG output (in isometric) of our algorithm viewed using Adobe SVG viewer 3.03 plug-in for Internet Explorer.

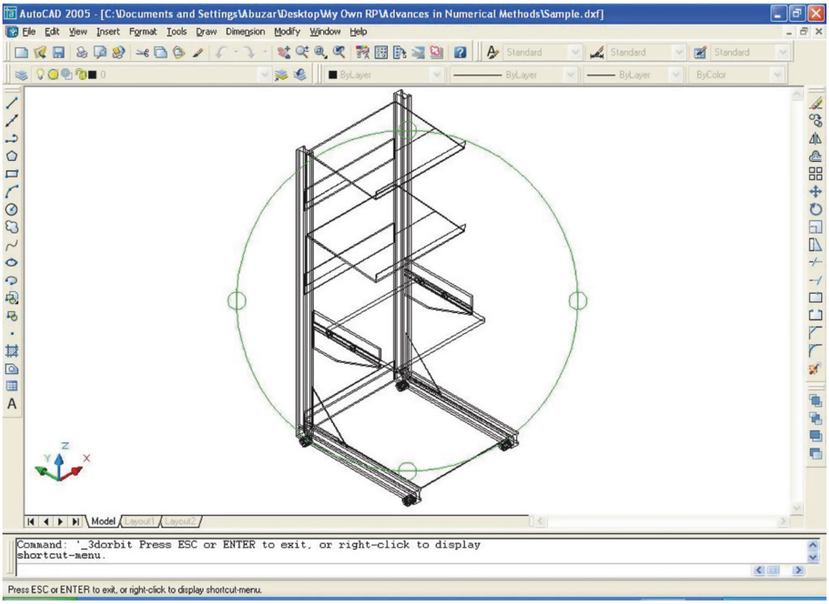
25.3 Comparison between SVG and DXF

We have compared both DXF and SVG formats on the basis of different parameters such as size of output file, time to load the document, support of 3D modeling, animation and rendering, user interactive editing, and the provision of different entities, layers, and dimensions.



Select any Bitmap image by clicking on the Browse Button

Fig. 25.2. Interface of the application


AutoCAD 2005 - [C:\Documents and Settings\Abuzar\Desktop\My Own RP\Advances in Numerical Methods\Sample.dxf]
File Edit View Insert Format Tools Draw Dimension Modify Window Help
Standard
Standard
Standard
ByLayer
ByLayer
ByLayer
ByColor
Model
Layout1
Layout2
Z
Y
X

```
Command: '_3dorbit Press ESC or ENTER to exit, or right-click to display
shortcut-menu.
```

Press ESC or ENTER to exit, or right-click to display shortcut-menu.

Fig. 25.3. DXF output in AutoCAD

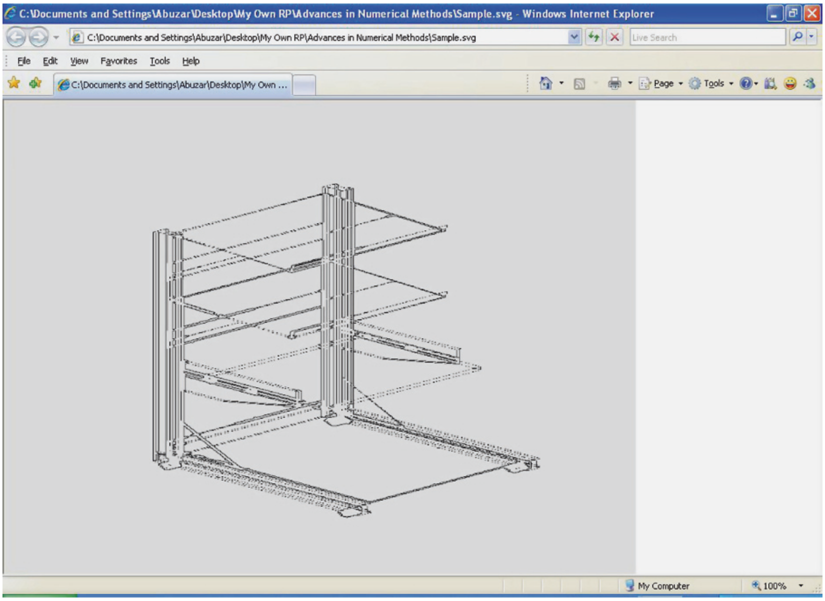

C:\Documents and Settings\Abuzar\Desktop\My Own RP\Advances in Numerical Methods\Sample.svg - Windows Internet Explorer
C:\Documents and Settings\Abuzar\Desktop\My Own RP\Advances in Numerical Methods\Sample.svg
Live Search
File Edit View Favorites Tools Help
C:\Documents and Settings\Abuzar\Desktop\My Own ...
Page
Tools
My Computer
100%

Fig. 25.4. SVG Output in Web browser

Our experiments showed that an SVG file is one-fourth of the DXF file in size, on average. A diagrammatic representation is shown in Fig. 25.5.

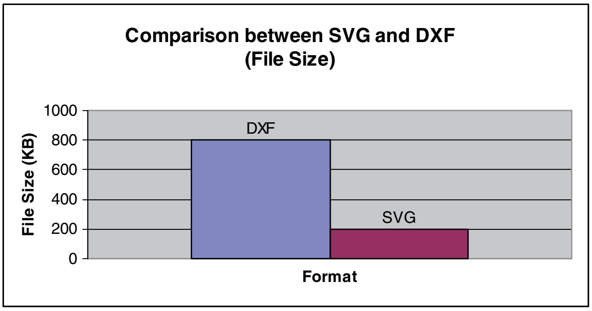


Fig. 25.5. Comparison between SVG and DXF w.r.t. file size

Our analysis resulted in another interesting fact about SVG format that its loading time is approximately one-sixth of DXF format loading time as shown in Fig. 25.6.

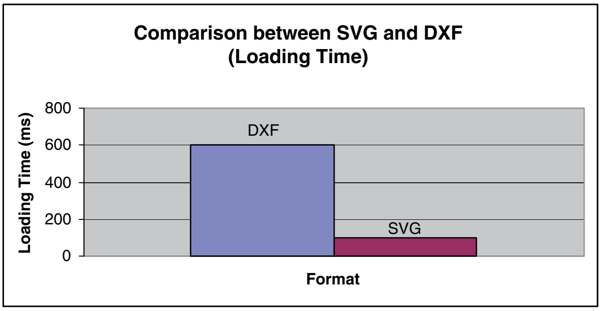


Fig. 25.6. Comparison between SVG and DXF w.r.t. loading time

It is worthy to mention that the loading time is not directly proportional to SVG file size; instead a logarithmic increase in time with increase in file size is observed and is shown in Fig. 25.7.

The above facts favor the suitability of SVG format to represent engineering drawings over World Wide Web.

There are some limitations of SVG, for example, it does not support 3D modeling implicitly; however, it does support rendering and animation. DXF format is capable of handling 3D features, layers, and dimensions of the line drawings. User interactive editing can be introduced in SVG documents with the help of CSVG [7–9]. Both formats support different entities with DXF being capable of handling 3D entities as well. Both types of documents can be rendered. The comparison is summarized in Table 25.1.

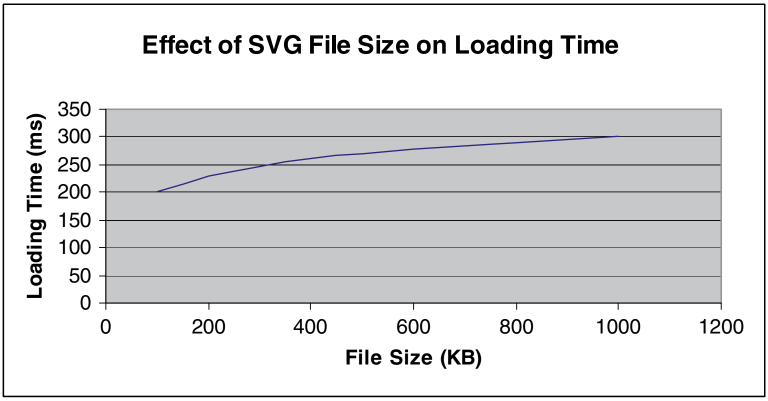


Fig. 25.7. Effect of SVG file size on loading time

Table 25.1. Comparison between SVG and DXF formats

Parameters	SVG	DXF
File size	Less (approximately one-fourth)	More
Loading time	Less (approximately one-sixth)	More
3D provision	No	Yes
Rendering	Yes	Yes
User interaction	Yes (with CSVG)	Yes

25.4 Summary

In this chapter, we have developed a technique to represent isometric engineering drawings in SVG format. Our technique can generate 3D DXF representation as well. We have concluded that SVG representations are more suitable for the distribution of engineering documents over World Wide Web after an exhaustive comparison on the basis of critical features of these formats. SVG representations are more appropriate from the point of view of file size and loading time.

25.5 Future recommendations

Our approach can be extended to 3D visualization of engineering drawings. Interested readers may consult [16, 17]. Another extension of our work may be the representation of different frames of a multipaged engineering document. Reading of [18] is recommended. Both these future recommendations can be implemented in a mobile-computing environment [19].

References

1. Quint A (2003) Scalable vector graphics. IEEE Multimedia 10(3):99–102
2. Fahiem MA, Haq SA, Sabir MR (2007) Comparison of 3D reconstruction techniques for engineering drawings from orthographic projections. 6th WSEAS International Conference on Applications of Electrical Engineering
3. Fahiem MA, Haq SA, Saleemi F (2007) A review of 3D reconstruction techniques from 2D orthographic line drawings. 2nd International Conference on Geometric Modeling and Imaging
4. Fahiem MA, Haq SA, Saleemi F, Tauseef H (2007) 3D reconstruction: Estimating depth of hole from 2D camera perspectives. European Computing Conference
5. Fahiem MA (2007) 3D reconstruction of solid models from 2D camera perspectives of engineering objects. European Computing Conference
6. Battiato S, Gallo G, Messina G (2004) SVG rendering of real images using data dependent triangulation. 20th Spring Conference on Computer Graphics
7. Battiato S, Barbera G, Blasi GD, Gallo G, Messina G (2005) Advanced SVG triangulation/polygonalization of digital images. IS&T/SPIE 16th Annual Symposium on Electronic Imaging, Science and Technology
8. Battiato S, Costanzo A, Blasi GD, Gallo G, Nicotra S (2005) SVG rendering by watershed decomposition. IS&T/SPIE 16th Annual Symposium on Electronic Imaging, Science and Technology
9. Battiato S, Blasi GD, Gallo, Nicotra S, Messina G (2005) SVG rendering of digital images: An overview. 13th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision
10. Battiato S, Blasi GD, Gallo G, Messina G, Nicotra S (2005) SVG rendering for internet imaging. 7th International Workshop on Computer Architecture on Machine Perception
11. Bardos GJ, Tirtowidjojo JJ, Marriott K, Meyer B, Portnoy W, Borning A (2001) A constraint extension to scalable vector graphics. 10th International Conference on World Wide Web
12. McCormack B, Marriott K, Meyer B (2004) Constraint SVG. 13th international World Wide Web conference on Alternate Track Papers & Posters
13. Mathis RM (2005) Constraint scalable vector graphics, accessibility and the semantic web. IEEE Southeast Conference
14. Molina F, Sweeney B, Willard T, Winter A (2007) Building cross-browser interfaces for digital libraries with scalable vector graphics (SVG). International Conference on Digital Libraries
15. Harit G, Garg R, Chaudhury S (2007) An integrated scheme for compression and interactive access to document images. International Conference on Computing: Theory and Applications
16. Ying J, Gracanin D, Lu CT (2004) Web visualization of geo-spatial data using SVG and VRML/X3D. 3rd International Conference on Image and Graphics
17. Yan W (2006) Integrating web 2D and 3D technologies for architectural visualization. 11th International Conference on 3D Web Technology

18. Christel MG, Huang CH (2001) SVG for navigating digital news video. 9th ACM International Conference on Multimedia
19. Su X, Chu CC, Prabhu BS, Gadh R (2006) Enabling engineering document in mobile computing environment. International Symposium on a World of Wireless, Mobile and Multimedia Networks

Chapter 26

Implementation of the box-counting method in radiographic images

K. Harrar,[1] L. Hamami[2]

Département d'Electronique, Laboratoire Signal et communications, Ecole Nationale Polytechnique, BP 182 El-Harrach, 16200 Alger, Algérie,
[1] hk_robot@yahoo.fr, [2] latifa.hamami@enp.edu.dz

Abstract. In recent years, there has been a growing interest in the physics and physical chemistry of fractals. Fractal systems show new physical properties and anomalous behavior. The box-counting method is an appropriate method of fractal dimension evaluation for images with or without self-similarity. The objective of this chapter is the introduction of the fractal theory and the fractal dimension in the radiographic images, and the implementation of the box-counting method for the segmentation of the images. However, this technique, including processing of the image and definition of the range of box sizes, requires a proper implementation to be effective in practice.

Keywords. Box-counting method, Fractal, Fractal dimension, Radiographic images, Side length, Self-similarity

26.1 Introduction

The name fractal was coined in 1975 by Benoit B. Mandelbrot, the founder of modern fractal science. It is derived from the Latin verb frangere: to break (fractum: broken), cf. words like fraction, fracture, fragment, etc.

A fractal is defined as a mathematical point set or a physical (chemical, biological, etc.) system with a fractional geometrical (spatial) dimension (more exactly: dimensionality). This dimension is called the fractal dimension

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_26,

or Mandelbrot dimension $\overline{d}$. Its rigorous and general mathematical foundation is given by the Hausdorff dimension d_{H} [1]. There also exist typical fractals which (by chance) carry integer dimensions, e.g., certain Viscek fractals or the Wiener trajectory of a Brownian particle during normal diffusion ($\overline{d} = 2$). This fact indicates difficulties to give a suitable definition of the fundamental concept "fractal" which is not too narrow and not too wide, these difficulties being shared by definitions of many other basic notions. Objects embedded in our common (Euclidean) space are characterized by (topological) dimensions $d = 1, 2$, or 3.

Historically, it is in the work of the mathematicians Cantor and Peano, at the end of the 19th century, where one finds the first references to sets often considered as pathological, whose geometry is particularly complex and is structured.

In 1919, Hausdorff proposed a new definition of the dimension of a set which can take non-integer values and which makes it possible to account for the degree of irregularity of these objects. One of the greatest merits of Mandelbrot is to have known that "the exception is often the rule" and to have shown that these fractured structures (singular) are in fact very present in nature. The profiles of our mountains, the various ramified geometries that the trees constitute, the rivers, or the overlaps of the bronchi in the lungs are as many examples as one can apprehend within the definite federator framework by Mandelbrot.

A fractal is a geometrical figure or a natural object which combines the following characteristics:

1. Its parts have the same form or structure that the whole, with this close which they are on a different scale and can be slightly deformed.
2. Its form, either extremely irregular, or extremely is stopped or split up with the remainder, in any scale of examination.
3. It contains "distinctive elements" whose scales are very fast varied and cover a large range.

26.2 The dimension and fractals

In order to understand the notion of a fractional dimension attributed to a physical system, let us consider some specific fractal structures. A good introductory example is a solid surface with defects like steps, dislocations, cavities with regard to its capacity to adsorb say atoms (Fig. 26.1). With increasing number of defects, the adsorptive capacity becomes larger and larger and approaches that of a sponge, to which can be assigned a dimension $d = 3$. Therefore, it makes sense to introduce an effective dimension $\bar{d}$ regarding adsorption, where $2 < \bar{d} < 3$. The topological dimension of the surface is now $d = 2$ as before: One indeed deals with a surface, even if curvature and many twists are present. The embedding dimension (or Euclidean dimension) of the system is $d_E = 3$.

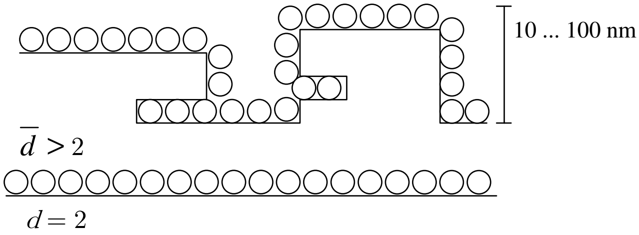


Fig. 26.1. A real solid surface shows a larger capacity to adsorb atoms than an ideal flat surface

The surface of a real solid is a typical stochastic fractal; In principle, one has to distinguish two types of fractals:

- Regular, deterministic or ideal fractals, model fractals; these fractals are constructed as mathematical objects according to a well-defined (unique) prescription or algorithm.
- Irregular, random or stochastic (or statistical) or real fractals; they are objects of nature or in the laboratory, yet they can be constructed theoretically as well.

The fractal systems of interest in physical science can be divided into three classes, typical representatives of which are

1. Natural objects (systems with mass) like solid surfaces, enzyme surfaces, irregular porous solids, e.g., catalysts, polymer systems, percolation networks, irregular particle aggregates.
2. Particle trajectories of classical Brownian motion (Wiener process), particle trajectories of anomalous diffusion, e.g., on fractal systems (the so-called fractional Brownian motion [2]) or due to time correlations in non-Markovian stochastic processes, quantum mechanical state vectors (wave function) of disordered electronic systems.
3. Sets in state diagrams, in particular boundaries of attractors and attractor basins of chaotic systems, i.e., classical dissipative dynamical systems with deterministic chaos, and sets of complex zeros of partition functions in the frame of statistical thermodynamics (Lee–Yang theorem).

In order to define and investigate the basic properties of these systems, it is convenient to consider ideal deterministic fractals. A typical example is the elementary Sierpiński fractal (Sierpiński network, lattice, triangle, or gasket); its construction is indicated in Fig. 26.2. With growing number n of construction steps the structure becomes finer and finer; nevertheless, one has nothing else than a surface with holes, i.e., an object with the geometrical (topological) dimension $d = 2$. Only in the limit $n \rightarrow \infty$ is a fractal generated; its dimension is $\overline{d} = \ln 3/\ln 2 = 1.5849 \ldots < 2$. The Sierpiński fractal is a prototype of a model for the physics of fractals and renders good services in the exploration of the new and sometimes strange phenomena occurring in fractal systems.

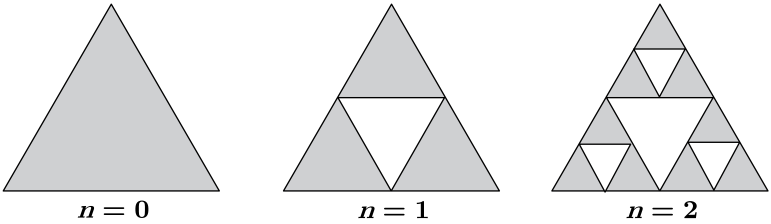


Fig. 26.2. Construction of a Sierpiński gasket; n: number of construction steps. Sketched are prefractals; the proper fractal is created in the limit $n \rightarrow\infty$

A known method to measure a length, a surface, or a volume consists in covering sets with paving stones (then it is only a question of counting the number of paving stones to obtain the length, surface, or volume of the set), of which the length, surface, or volume is taken as measuring unit.

If ε is the side (length standard) of a paving stone, measurement obtained is [1]

$$M = N \,.\, \varepsilon^{\mathrm{d}} = N\mu \qquad (26.1)$$

where μ is the unit of measure (length, surface, or volume).

Mandelbrot postulates that there are curves of intermediate size between 1 and 2 of surfaces of size higher than 2 and that these objects precisely have the property to have no length or a precise surface, not more than one volume does not have a surface or a square does not have a length. This dimension, intermediary between the integer values, was baptized neologism “fractal” so that no confusion is made between a traditional surface (of $D = 2$ dimension).

One is brought to believe that a geometrical object, about scale, can also generate the small as well as the big details. Such an object will be known to have an internal homothety or to be self-similar. It is known that if one transforms a line by a homothety of arbitrary ratio, whose center belongs to it, one finds this same line, and it is the same for any plane and entire Euclidean space. One can generalize even for non-integer dimensions, as this definition indicates it.

If a fractal object S is divided in N objects similar to S in a homothety $1/r$, the dimension of homothety or fractal dimension is the ratio

$$D = \frac{\log N}{\log 1/r} \tag{26.2}$$

It should be checked that the curves do not have double points. It is not the same with other curves which have a double infinity of points. It follows that for them, the concept of paving changes significance and that the definition of the homothety dimension becomes debatable.

26.3 The box-counting method

Linear fractal images are the outcome of absolute generating processes and the information related to each step of the process can be calculated exactly [2, 3]. For instance, in a linear fractal image like Koch boxes, seven new segments, three times smaller than the previous segment, are generated at each step in the generating process. Therefore, an exact mathematical calculation procedure follows. After the first step, the image contains seven segments whose size is 1/3 that of the initial value; after the second step, the image contains 49 segments of size 1/9; and so on. In linear fractals, even after two steps of the generating process, the fractal dimension can be calculated exactly.

In non-linear fractal images, because of the existence of random elements there is a statistical (i.e., not deterministically mathematical) generating process and the information available at each step of the generating process is not exact [4]. Natural fractals fall into this category of fractals. For this reason, an appropriate method is needed to estimate the fractal dimension of non-linear images.

Among the techniques discussed by Mandelbrot [5], the box-counting method is found the most adapted for the estimate of fractal dimension [6]. Voss, Keller and Sarkar [7–9] carried out a box-counting method, the purpose of which is to consider the average number, noted $N\ (r)$, of cubic boxes with fixed side length r, necessary to cover the image, considered as a surface in R^3 space. For that we estimate $P\ (m, r)$, the probability that one box of size r, centered on an arbitrary point of surface, contains m points of the set. We have thus [10]

$$\forall r, \sum_{m=1}^{Np} P(m,r) = 1 \tag{26.3}$$

where *Np* is the number of possible points in the cube.

The estimate of the average number of disjoined boxes to cover surface is

$$N(r) = \sum_{m=1}^{Np} N(m,r) = \sum_{m=1}^{Np} \frac{P(m,r)}{m} \tag{26.4}$$

The estimate by the least squares method of the slope of the group of dots (log (*r*), − log (*N* (*r*))), obtained with boxes of increasing size *r*, gives the estimate of fractal dimension. The Algorithm 26.1 [10] presents this calculation:

```
Initialization:
FOR r =1 to $r_{max}$ and m = 1 to $r^3$ DO
     P(m,r) = 0
FOR any site s of the image DO
     BEGIN
     For r = 1 to rmax DO
        BEGIN
          - Center a cube with dimension r on [s,A[s]]
          - Count the number m of pixels of the image
             which belong to this cube
          -  Increment P(m,r) by 1
        END
     END
FOR  r = 1 to rmax DO
          $N(r)=\sum_{m=1}^{Np}\frac{P(m,r)}{m}$
Estimate by the method of least squares the slope D of the
curve (log (r), -log (N(r)))
```

Algorithm. 26.1. The box-counting method algorithm

Recent work on the fractal dimension using the box-counting method in the radiographic images was done by Imai et al. [11], who have conducted a fractal analysis (with the box-counting method for a binary images) of low-dose digital chest phantom radiographs and calculated fractal-feature distance using the fractal dimension. This method uses a lot more materials and methods than it offers.

Podsiadlo et al. [12] have also studied differences in trabecular bone texture between knees with and without radiographic osteoarthritis detected by fractal methods. A newly developed augmented Hurst orientation transform (HOT) method was used to calculate texture parameters for regions selected in X-ray images of non-OA and OA tibial bones. This method produces a mean value of fractal dimensions, FDs in the vertical ($\mathrm{FD_V}$) and horizontal ($\mathrm{FD_H}$) directions and along a direction of the roughest part of the tibial bone ($\mathrm{FD_{Sta}}$), fractal signatures, and a texture aspect ratio (Str).

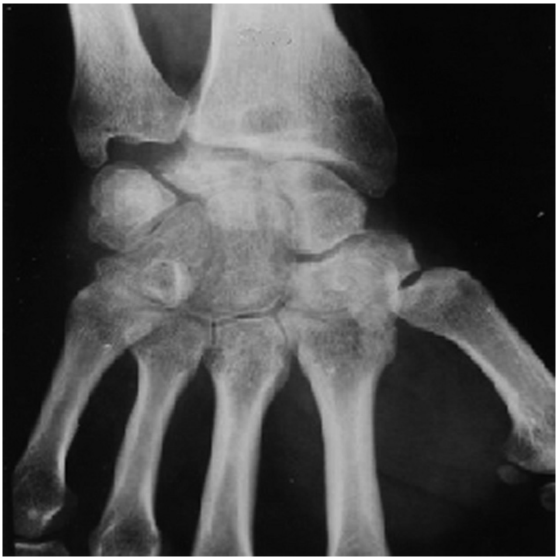


Fig. 26.3. Radiographic image of the hand

In what follows we will apply the method for the image X-ray test (Fig. 26.3) of size 256 × 256 pixels in gray scale and see the results of the calculation of fractal dimension. For that we propose to plot curves giving the number of boxes versus their side length *r*, then the straight regression line which estimates as well the possible log (*N*(*r*)) versus log (1/*r*); the plot will be done on all the points and outdistance between the size of boxes is equal to 1.

Figure 26.4 represents the number of boxes according to their sizes (*r*), in this case $r = 1 – 50$. We notice that the smaller the *r,* the more the number of boxes, and inversely, the more we increase *r,* the smaller the number of boxes. In addition, we see a sharp fall of number of boxes after length $r = 1$. When *r* reaches 40 or beyond, it becomes increasingly challenging to count the number of boxes, for this we give a table of some values of *r* and the number of the boxes (Table 26.1).

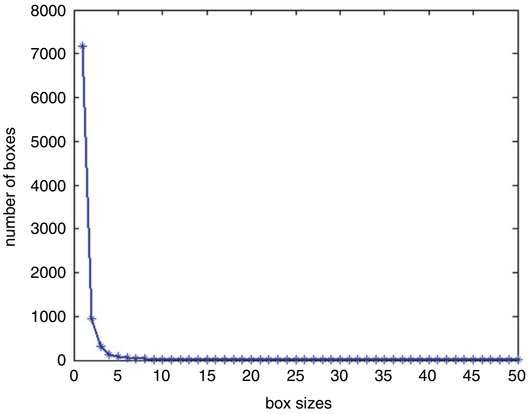


Fig. 26.4. The plot of the number of boxes versus the side length *r* for the radiographic image

Table 26.1. Some values of the number of boxes according to the side length *r*

r	*N*(*r*)
1	7181
2	951
3	136
5	74
10	20
15	6
30	3
40	2
50	1

r side length of the boxes, N(r) number of the boxes

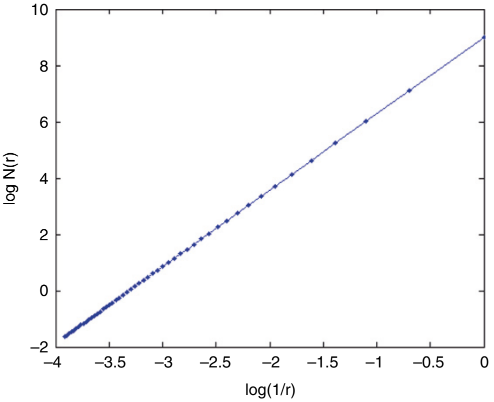


Fig. 26.5. The plot regression of the curve log *N*(*r*) versus log (*1*/*r*) by the method of least squares

In this experiment, we obtained a dimension D = 2.72 for r = 1–50, the plot of regression is illustrated in Fig. 26.5. The only parameter which can influence the calculation of fractal dimension is the actual range of the sizes of the windows (boxes); this point will be expanded in the next section.

26.4 The range of box sizes

In using the box-counting method, challenges arise when the range of box sizes is to be determined. In particular, defining the largest and smallest box sizes to use requires extreme care. In addition, the positioning of the grid to superimpose on the image has a critical effect on resulting estimate of fractal dimension. Therefore, both factors should be verified [13, 14].

To find fractal dimension, it is challenging to choose the sizes of the boxes for complex images like the radiographic images. Several tests were carried out on different sizes and the values of fractal dimension were obtained.

The change in range is illustrated in Fig. 26.6, followed by the plot of regression (Fig. 26.7), and we notice that the number of boxes changes.

The change in range resulted in different results for the same images compared to the previous range, the fractal dimension in this case is D = 1.84 and for this $r_{\min}$ = 2 and $r_{\max}$ = 40. Therefore, we can conclude that there are four possibilities:

$r_{\min}$ (small, large), $r_{\max}$ (small, large)

Of course, it is necessary that they are sufficiently isolated to avoid the effect of overlapping.

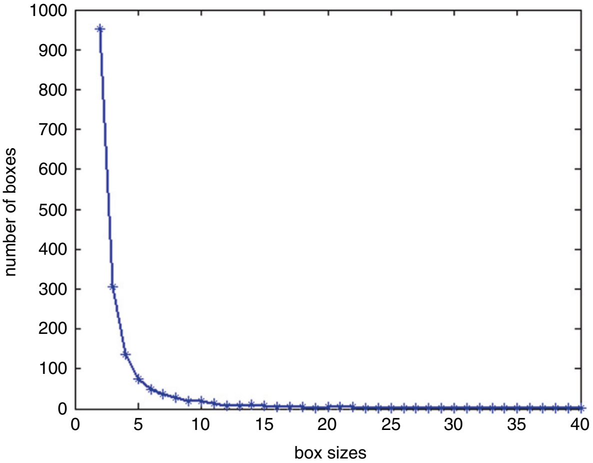


Fig. 26.6. The box-counting method with $r_{\min}$ =2, $r_{\max}$ = 40

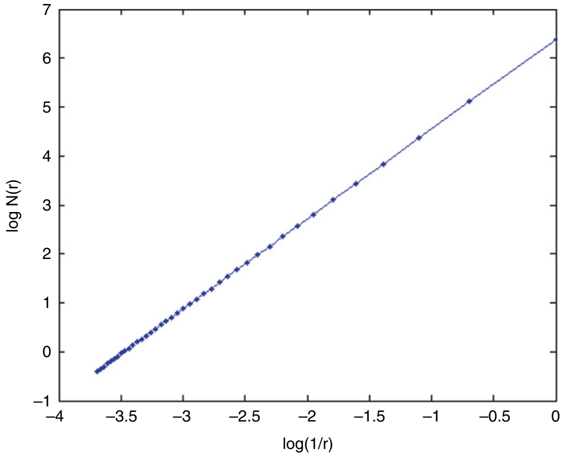


Fig. 26.7. The regression of the curve log $N(r)$ by the least squares method

We give in Table 26.2 certain values of fractal dimension corresponding to the range of box sizes:

Table 26.2. Some examples of the fractal dimension

$\mathrm{r}_{\min}$ – $\mathrm{r}_{\max}$	D
1–50	2.72
1–40	2.66
2–40	1.84
2–50	2.03
3–50	1.48

$r_{\min}$–$r_{\max}$ range of box sizes, D fractal dimension.

26.5 Choice of box sizes

If we see Table 26.2, we notice a variability of the values for the fractal dimension; an improvement of the original method can be implemented by the choice of $r_{\min}$ and $r_{\max}$.

Up to now, it is difficult to choose the side length r of the boxes considering the complexity of the images, but we can always take a compromise between $r_{\min}$ and $r_{\max}$. These two parameters are not only influential to one another, but more importantly on the calculation of the fractal dimension.

For the choice of r_{min} and after several tests carried out, it is safe to say that we can choose r from 2; this will give us a greater probability of finding at least one box; so, for $r_{\text{min}} = 1$, we are confronted with the problem of the pixel size, where a 1×1 box cannot be centered due to its single pixel point. The isolated pixels at $r_{\text{min}} = 1$ should not be taken into consideration, so $r_{\text{min}} = 2$ is a good choice to find a box and to detect pixels going up to 4 (2×2).

For r_{max}, it is fixed in the following way: As soon as $N(r)$ get too close or overlap each other, we stop the process; from there, an extraction of the maximum value of r is carried out. We can also fix it from the moment when the number of windows (boxes) decreases toward 0 (this number must stop at 1); from this moment it is not necessary to repeat the iterations for larger sizes, it is a waste of time, from the optimization of the computing time. In addition, the value of r_{max} should not exceed the framework of the image, where

- If the horizontal size of the image (x) ≤ the vertical size of the image: $r_{\text{max}} < \text{x}/2$.
- If the vertical size of the image (y) ≤ the horizontal size of the image: $r_{\text{max}} < \text{y}/2$.

For these reasons the choice of, the range, the beginning, and the end of the side length r are very important to calculate the fractal dimension in any image.

26.6 Determination of breakpoint in the log–log plot

We considered two procedures to verify the breakpoint between points used to fit a straight line in the first part of the log–log plot and the rest of the points in the plot. The first procedure consists in fitting a regression line over all points and then eliminating points one by one, starting from the smallest box size to the largest box size, until a regression line with almost all points attached is obtained. The last point on this regression line will be the breakpoint that defines the best smallest size for the range of box sizes. In this procedure, more sizes may sometimes be needed. For instance, using this procedure for data of Fig. 26.8, a few more box sizes between the fifth and sixth were indicated since the regression lines fitted to the first five and first six points were attaching almost all the points, with little deviation for the first six points.

Another procedure that we considered is based on dummy variables used in a regression analysis with two regression lines [15]. This procedure provided identical results for the data of Fig. 26.8. Since it is expected that the log–log plot cannot satisfy two straight lines for some images, this procedure must be applied with caution.

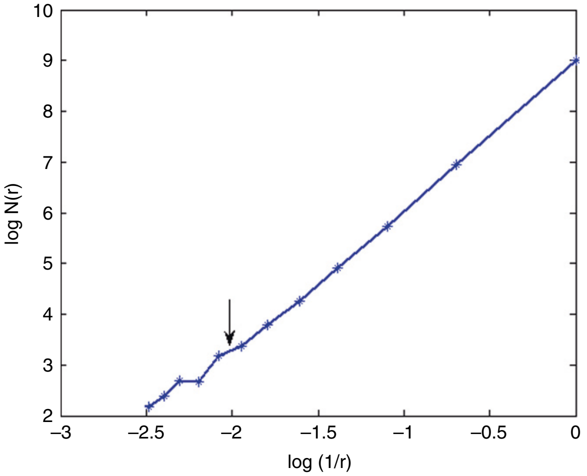


Fig. 26.8. Illustration of the breakpoint

26.7 Conclusion

Throughout this chapter, we studied the fractal dimension for the radiographic images with two dimensions; after an introduction on the theory of the fractals and its dimension, we discussed that these images were complex and non-linear. For the calculation of their fractal dimension it would have been necessary to find a method that can be freed from the existence of the phenomenon of the statistical random elements that these images present. For that the box-counting method is up to the task and gave good results.

We did not only calculate the fractal dimension but also noticed that the range of the box sizes influenced much of the calculation of this dimension, for that we presented some calculations of fractal dimension corresponding to a certain range of *r*; however, extreme care must be taken in the choice of this range for the analysis.

The choice of size for these boxes is worked out for better calculation of the fractal dimension.

The importance and usefulness of the fractal concept can be summarized as follows:

1. Fractals possess mathematically strange properties and imply very interesting novel physical phenomena.
2. Fractals are suitable to model complicated real systems.
3. Fractals allow for the application of rigorous scaling and renormalization methods without approximations.

References

1. Gouyet JF (1992) Physique et structures fractales. Masson, Paris
2. Sokolov IM, Klafter J, Blumen A (2002) Physics Today 55:48
3. Pfeifer P (1985) Chimia. 39:120
4. Maître H (2003) Le traitement des images. Lavoisier, Paris
5. Mandelbrot BB (1983) The fractal geometry of nature. Freeman, New York
6. Fouroutan-pour K, Dutilleul P, Smith DL (1999) Advances in the implementation of the box-counting method of fractal dimension estimation. Applied Mathematics and Computation 105:195–210
7. Voss R (1986) Random fractals: characterization and measurement, in Scaling phenomena and disordered systems. In: Pynn R, Skjeltorp A (eds) Plenum Press, New York, pp 1–11
8. Keller JM, Chen S, Crownover RM (1989) Texture description and segmentation through fractal geometry. CVGIP 45:150–166
9. Sarkar N, Chaudhuri BB (1992) An efficient approach to estimate fractal dimension of textural images. Pattern Recognition 25(9):1035–1041
10. Cocquerez JP, Philipp S (1995) Analyse d'images: Filtrage et segmentation. Masson Paris
11. Imai K, Ikeda M, Enchi Y, Niimi T (2007) Fractal-Feature Distance Analysis of Radiographic Image. Academic Radiology 14(2):37–143
12. Podsiadlo P, Dahl L, Englund M, Lohmander LS and Stachowiak GW, (2007) Differences in trabecular bone texture between knees with and without radiographic osteoarthritis detected by fractal methods. Osteoarthritis and Cartilage (in press)
13. Kaye BH (1994) A Random walk thrugh. Fractal Dimensions, Weinheim, New York
14. Soddel J, Seviour R (1994) Using box counting techniques for measuring shape of colonies of filamentous micro-organisms. In: Stonier RJ, Yu XH (eds) Complex systems: Mechanism of adaptation, IOS Press, Amsterdam
15. Draper NR, Smith H (1981) Applied regression analysis. Wiley, New York

Chapter 27

Aeroelastic simulation of wind turbine blades

Z.L. Mahri, M.S. Rouabah, Z. Said

Département de Génie Climatique, University of Mentouri (Constantine), Alegria

Abstract. The aim of this chapter is to compute dynamic stresses acting on wind turbine blades. These stresses are essential in predicting fatigue of the rotor.
The turbine rotor is exposed to wind loading of a cyclic nature, making it vulnerable to cumulative fatigue damage.
The approach used in this chapter is based on the analysis of the blade movement, by solving the blade motion equation, in order to obtain dynamic stresses.
In the first part of this chapter, modal analysis of the blades is carried out in order to compute the frequencies and the mode shapes. The results obtained in the modal analysis are used to compute dynamic stresses, for different wind loads, at the root region of the blades.

Keywords. Wind energy, Structural dynamics, Aerodynamics, Numerical analysis

27.1 Introduction

The prediction of the dynamic behavior constitutes one of the most important processes in the design of wind turbines, because it can be useful in estimating the energetic performance of the wind turbine as well as in predicting fatigue and structural problems of this machine. The study of this dynamic behavior can be undertaken by various analysis methods [1].

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_27,

The aim of dynamic analysis is to compute dynamic loads and stresses; these stresses are essential in predicting fatigue.

This analysis has a particular importance knowing that the major cause of wind turbine failure is fatigue. The turbine blades are exposed to wind loading of a cyclic nature making them vulnerable to cumulative fatigue damage**.**

The fatigue estimation of the blades is useful in preventing blade breakage; however, the difficulty in predicting fatigue is in large part due to an insufficient knowledge of the dynamic behavior of these machines [2].

The dynamic analysis for a rotating blade must go through several steps of calculations: among them are the calculation of mode shapes and frequencies and the computation of displacements and stresses acting on the blades.

The approach used in this work is based on the study of the blade movement, by solving blade motion equation, in order to obtain the dynamic stresses. The aerodynamic consideration implies that load calculation is dependent on the shape (the form) of the blade; on the other hand, these loads will cause deformation of the blade.

This interdependence between the aerodynamic load and the shape of the blade is a special characteristic of aeroelasticity problems. This factor complicates the solution of the problem.

Once the dynamic stresses are calculated, the fatigue of the blades can be estimated using the cumulative damage theory [3].

27.2 Calculation of bending frequencies

The equilibrium equations for a blade element having a length ℓ are given as follows [4] (Fig.27.1):

$$G_{n+1} = G_n + \int_{x_{n+1}}^{x_n} \Omega^2 x \ m \ dx = G_n + \frac{1}{2}\Omega^2 m (x_n^2 - x_{n+1}^2) \tag{27.1}$$

$$V_{n+1} = V_n - \int_{n+1}^{n} \omega^2 Z \ m \ dx = V_n - \frac{1}{2} m \omega^2 \ell (Z_n + Z_{n+1}) \tag{27.2}$$

$$M_{n+1} = M_n - G_n (Z_n - Z_{n+1}) - V_n \ell + \int_{x_{n+1}}^{x_n} (x - x_{n+1}) \omega^2 Z \ m \ dx - \tag{27.3}$$

$$(Z - Z_{n+1}) \Omega^2 x \ m \ dx$$

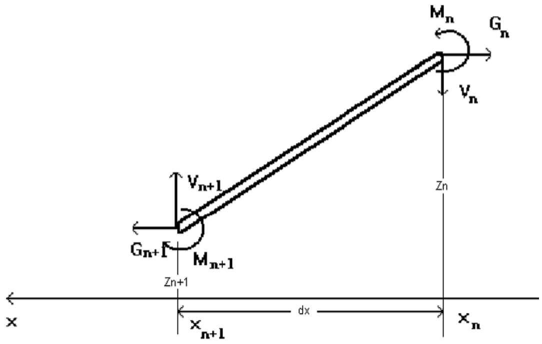


Fig. 27.1. Loads applied on a blade element

where
V_n the shear force at the node *n, M_n* the bending moment at node *n, G_n* the centrifugal force at the node *n, Z_n* the deflection of the node *n, M* the linear mass, and ω the natural frequencies.

After assembling all the blade elements the sequences *V*, *G*, and *M* can be transformed to a matrix equation of the form

$$(A - I/\omega^2)Z = 0 \tag{27.4}$$

The natural frequencies ω are determined from the eigenvalues of the matrix ***A***.

27.3 Resolution of the flapwise motion equation

In this section, the dynamic stresses are calculated using a numerical approach to solve the blade motion equation, where the blade is considered as a continuous system.

The blade motion equation in the flapwise direction can be expressed by the following equation [4]:

$$\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 Z}{\partial x^2}\right) - \frac{\partial}{\partial x}\left(G\frac{\partial Z}{\partial x}\right) + m\frac{\partial^2 Z}{\partial t^2} = \frac{\partial F}{\partial x} \tag{27.5}$$

where $G = \int_{x}^{L} m\,\Omega^2\, x\, dx$, t is time, and F is the aerodynamic load.

27.3.1 Resolution of the free vibration equation (calculation of the blade mode shapes)

The mode shapes can be calculated in case of free vibrations (if the blade is not exposed to external loads) using Eq. (27.5), which can be written as:

$$\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 Z}{\partial x^2}\right) - \frac{\partial}{\partial x}\left(G\frac{\partial Z}{\partial x}\right) + m\frac{\partial^2 Z}{\partial t^2} = 0 \tag{27.6}$$

By taking $Z = S(x).\varphi(t)$ and using the method of variable separation, a set of two ordinary differential equations is obtained:

$$\frac{d^2}{dx^2}\left(EI\frac{d^2 S}{dx^2}\right) - \frac{d}{dx}\left(G\ \frac{dS}{dx}\right) - m\ \omega^2 S\ =\ 0 \tag{27.7}$$

$$\frac{d^2\varphi}{dt^2} + \omega^2\varphi\ = 0 \tag{27.8}$$

The values of the frequencies ω are taken from the preceding method. The boundary conditions of Eq. (27.7) are the following:

- At the fixed end of the blade:

$$\text{Displacement} = 0 \quad \Rightarrow \quad S(0) = 0 \tag{27.9}$$

$$\text{Slope} = 0 \quad \Rightarrow \quad \frac{dS(0)}{dx} = 0 \tag{27.10}$$

- At the free end of the blade:

$$\text{Bending moment} = 0 \quad \Rightarrow \quad \frac{d^2S(L)}{dx^2} = 0 \tag{27.11}$$

$$\text{Shear force} = 0 \quad \Rightarrow \quad \frac{dS^3(L)}{dx^3} = 0 \tag{27.12}$$

The numerical solution of Eq. (27.7) is complicated by its special boundary problem, having two initial values, at the fixed end (Eqs. (27.9) and (27.10)), and two final values, at the free end (Eqs. (27.11) and (27.12)).

In order to start a numerical solution of Eq. (27.7), the initial values $\frac{d^2 s(0)}{dx^2}$ and $\frac{d^3 s(0)}{dx^3}$ must be known. If a first guess of these values is made, a solution S(x) can be obtained using a numerical technique, but in this case the solution obtained does not necessarily satisfy the given boundary conditions. It is obvious that different initial values will give different solutions.

It has been verified that the predictor corrector method (Adam's formula) [5] can solve Eq. (27.7) with a good convergence, while the Runge–Kutta method has failed to reach convergence [6].

Taking $x_1 = \frac{d^2 S(0)}{dx^2}$ and $\mathrm{x}_2 = \frac{dS^3(0)}{dx^3}$

also

$$f\,(x_1,x_2) = \frac{d^2 S(L)}{dx^2} \tag{27.13}$$

$$g(x_1,x_2) = \frac{dS^3(L)}{dx^3} \tag{27.14}$$

This boundary condition problem can then be formulated as a problem of solving a set of two equations [7]:

$$f\,(x_1,x_2) = 0 \tag{27.15}$$

$$g(x_1,x_2) = 0 \tag{27.16}$$

Since the analytical expressions of the functions ***f*** and ***g*** are unknown but their numerical values can be obtained by the predictor corrector method, the secant method is used to obtain x_1 from Eq. (27.13) and x_2 from Eq. (27.14). This iterative procedure is repeated using the new values of x_1 and x_2 until convergence is reached.

This approach is advantageous in the calculation of the mode functions, since a larger number of points can be determined.

A Fortran computer program was implemented to perform these computations. The mode shapes obtained, using this approach, are represented by Figs. 27.2, 27.3, and 27.4

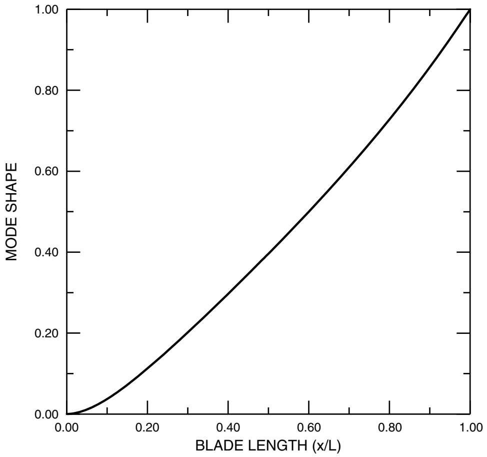


Fig. 27.2. First flapwise mode shape

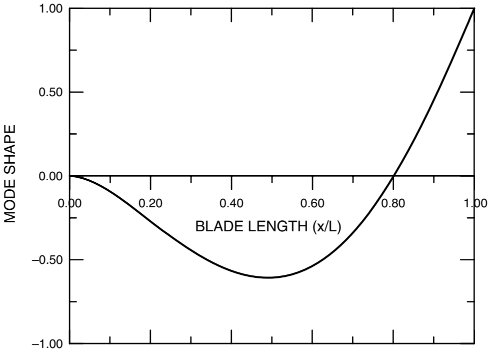


Fig. 27.3. Second flapwise mode shape

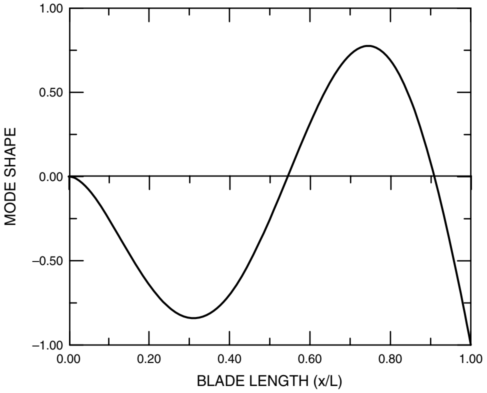


Fig. 27.4. Third flapwise mode shape

27.3.2 Resolution of the forced vibration equation (calculation of the blade dynamic stresses)

The forced vibration of a blade subjected to aerodynamic load is expressed by Eq. (27.5).

The solution ***Z*** (the displacement) can be expressed as follows:

$$Z=\sum_{i=1}^{n} S_i(x).\varphi_i(t) \tag{27.17}$$

where S_i is mode shape *i* and n is the number of modes.

These mode shapes verify the orthogonal property defined as [4]

$$\int_0^L m\,S_i(x)S_j(x).dx = 0 \qquad \text{if} \quad i{\neq}j$$

$$\int_0^L m\,S_i(x)S_j(x).dx = f(i) \qquad \text{if} \quad i{=}j$$

In order to solve Eq. (27.5) the response functions $\varphi_i(t)$ must be determined. If the expression of **Z,** given by Eq. (27.17), is substituted into Eq. (27.5), the following expression is obtained:

$$\sum_{i=1}^{n}\varphi_i(\mathrm{t})\left[\frac{d^2}{dx^2}\left(EI\frac{d^2S_i(\mathrm{x})}{dx^2}\right)-\frac{d}{dx}\left(G\ \frac{dS_i(\mathrm{x})}{dx}\ \right)\ \right]+m\sum_{i=1}^{n}\mathrm{S}_i(\mathrm{x}).\frac{\partial^2\varphi_i(\mathrm{t})}{\partial t^2}=\frac{\partial F}{\partial x} \tag{27.18}$$

According to Eq. (27.7) the expression between brackets can be replaced by $m\omega^2S_i(x)$, then Eq. (27.18) becomes

$$\sum_{i=1}^{n}\left(\frac{d^2\varphi_i(t)}{d\mathrm{t}^2}+\ \omega^2\varphi_i(t)\right)S_i(\mathrm{x})\ =\ \frac{1}{m}\frac{\partial F}{\partial x} \tag{27.19}$$

Equation (27.19) is multiplied by $mS_i(x)$ and then integrated from 0 to *L* with respect to *x*. Taking the mode orthogonally into account, one obtains

$$\frac{d^2\varphi_i(t)}{d\mathrm{t}^2}+\ \omega^2\varphi_i(t)=\frac{1}{\mathrm{f(i)}}\int_0^L\frac{\partial F}{\partial x}S_i(\mathrm{x})dx \tag{27.20}$$

By solving Eq. (27.20), the response functions $\varphi_i(t)$ and thereafter the displacements ***Z*** can be determined.

The previous simplification appears to succeed in separating the modes since Eq. (27.20) contains only one mode; however, the right member of this equation includes the aerodynamic loads ***F***, which depend on the shape of the blade (the displacement ***Z***) and thereby of all the modes. Thus, each mode is dependent on the rest of the modes.

This interdependence between the load and the shape of the blade (deformation) will complicate the solution of Eq. (27.20).

To start the resolution of Eq. (27.20), one should assume an initial blade form, such as a simple rigid (linear) deformation. Then the right member of this equation is computed for each mode. Then Eq. (27.20) becomes

$$\frac{d^2\varphi_i(t)}{dt^2} + \omega^2\varphi_i(t) = C_i \tag{27.21}$$

where C_i is the value of the right member calculated for the *ith* mode.

Each value C_i will be supposed constant, for a small time interval Δt. The solution of Eq. (27.21), in this case, will be

$$\varphi = \frac{C_i}{\omega^2}(1-\cos\omega t) + \varphi_0\cos\omega t + \frac{\varphi'_0}{\omega}\sin\omega t \tag{27.22}$$

where φ_0 and φ'_0 are the initial values.

The new modes calculated, using Eq. (27.22), are used to determine the displacements **Z** (a new form of the blade), which allows the calculation of the aerodynamic load and thus the new values C_i. This approach is repeated using the following iterative formulas derived from the solution (27.22):

$$\varphi_{j+1} = \frac{C_i}{\omega^2}(1-\cos\omega\Delta t) + \varphi_j\cos\omega\Delta t + \frac{\varphi'_j}{\omega}\sin\omega\Delta t \tag{27.23}$$

$$\varphi'_{j+1} = \frac{C_i}{\omega}\sin\omega\Delta t - \omega\varphi_j\sin\omega\Delta t + \varphi'_j\cos\omega\Delta t \tag{27.24}$$

It should be noted that the subscript *j*, in this case, represents the calculation step number.

At the beginning, this calculation procedure is repeated, until convergence is reached (in order to correct the initial form of the blade).

Afterward, this calculation is carried out for each sequential step of time [7].

The torsion motion equation is solved in a similar manner as the flapwise equation.

27.4 Equation of coupled movement

The two equations (bending–torsion) must be coupled in order to determine stresses and displacements. This is done using the equation of Brooks [4], which has the form

$$\frac{\partial^2}{\partial x^2}\left(EI\frac{\partial^2 Z}{\partial x^2}\right)-\frac{\partial}{\partial x}\left(G\frac{\partial Z}{\partial x}\right)+m\frac{\partial^2 Z}{\partial t^2}+Ft\left(\mathrm{x}\ ,\theta,\frac{\partial^2 \theta}{\partial t^2}\ \right)=\frac{\partial F}{\partial x} \tag{27.25}$$

where Ft is a known function.

In this equation, the effect of torsion on bending is taken into account. The particular difficulty encountered in the resolution of the coupled equation is due to the fact that the wind load depends upon the shape of the blade (deflection), since this load is a function of the incidence angle; on the other hand, the load deforms the shape of this blade. This interdependence between the aerodynamic load and the blade deflection is a source of nonlinearity that complicates the numerical resolution.

To solve this equation, an initial deflection is supposed and then an iterative method is used to correct this shape.

Results: Figure 27.5 represents the maximum normal stress and Figs. 27.5 and 27.6 represents the maximum shear stress for a wind speed of 3 m/s.

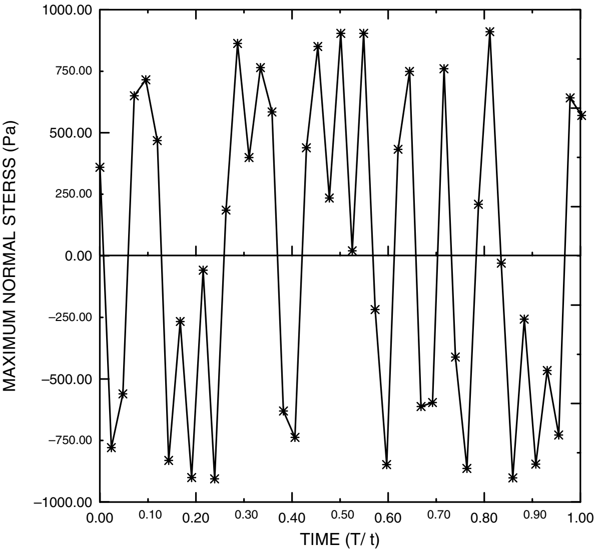


Fig. 27.5. Maximum normal stress at the fixed end

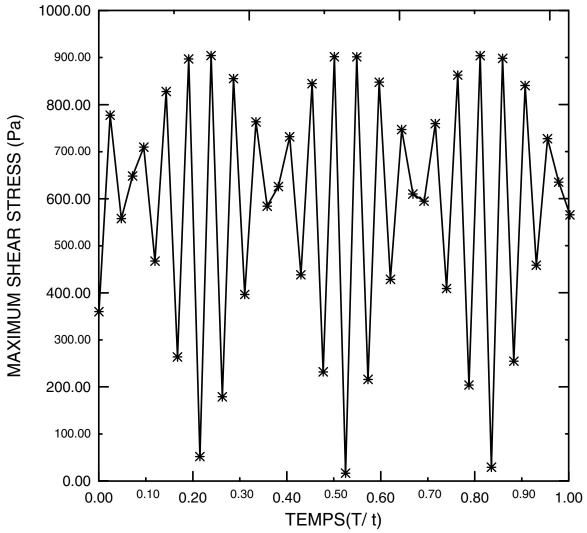


Fig. 27.6. Maximum shear stress at the fixed end

27.5 Conclusion

The dynamic stresses are calculated using a numerical approach to solve the blade motion equation, where the blade is considered as a continuous system.

The interdependence between the aerodynamic load and the blade deflection is a source of nonlinearity that complicates the numerical resolution.

The dynamic stresses calculated may be used to estimate the fatigue of the blade by the cumulative damage theory [8].

This fatigue estimation can be helpful in preventing blade breakage, which is a very frequent problem for wind turbines.

References

1. Younsi R (2001) Dynamic study of wind turbine blade with horizontal axis. European Journal of Machanics A: Solids 20:241–252
2. Walker JF (1997) Wind Energy Technology. John Wiley & sons, New York
3. Ronoldk KO (1999) Reliability-based fatigue design of wind turbine rotor blades. Engineering Structures 21:1101–1114

4. Bramwell A (1989) Helicopter Dynamics. Edward Arnold, London
5. Nougier J (1996) Méthode de calcul numérique. Masson, Paris
6. Mahri ZL (1998) résolution de l'équation des modes propres d'une pale d'éolienne. Conférence internationale sur les mathématiques appliquées et les sciences de l'ingénieur, Casablanca, Morocco
7. Mahri ZL (2002) Fatigue estimation for a rotating blade of a wind turbine. Revue des Energie Renouvelables UNESCO–CDER 5:39–47
8. Mahri ZL (2006) Calculation of dynamic stresses acting on the wind turbine blades. World Renewable Energy Congress IX, Florence, Italy, Proceeding ed Elsevier
9. Mahri ZL (1999) Fatigue estimation of a rotating blade. World Renewable Energy Congress, Perth, Australia

Chapter 28

Time-delay telerobot system control model research

Jin He,[1] Qingxin Meng[2]

[1]Department of Physics & Electrical Engineering, Yunnan Nationalities University, 12.1 Street, Kunmin 650031, P. R. China, Kmhjj12@163.com
[2]Mechanical and Electrical Engineering, Harbin Industry University, Harbin 150001, P. R. China, Hjmwh@sina.com

Abstract. This chapter introduces a new optimization of the event-based control model for the time-delay problem. With the waiting-time defect in traditional event-based teleoperation under consideration, this chapter presents the simulated force directly to the teleoperator. The buffer sequence mechanism is accepted at the telerobot end. This method ensures command execution in the sequencing event. Experimental results show that the control effect is satisfactory.

The main direction of bilateral research in teleoperation systems includes the studyof manipulation force and the sense of touch as measured by the robot sensor, etc. Because of the time-delay influence, it proposes an extremely incisive question to the control method: Namely, the existence of a time delay causes the system's performance to decrease, makes the entire teleoperation system unstable, and makes synchronism difficult. At the same time, the time delay has a variable time characteristic, so research on a time-delay robot control model is extremely necessary.

Keywords. Teleoperation, Time delay, Manipulator, Event-based control model, Control system

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_28,

28.1 Introduction

With the exploration of space and the deep sea, and other faraway places, teleoperation research has become more important than ever before. With the progress in accelerating computer networks and building new technology, and with the convenience of the Internet, research into robot teleoperations is becoming much more essential. In the meantime, forms of teleoperaton robotics have been used in many areas, including deep sea exploration, telemedicine, and so on. With the application requirement increment of Internet-based teleoperation, such as space exploration, hazardous environment working, researches, entertainment and education. Based on the research from current teleoperation techniques, this chapter discusses key technologies, such as bilateral, Internet properties, and time delay.

28.2 Control model

By analyzing existing telerobot system control models based on the variable-wave method or the scattering-theory control method, we concluded that the rationale of the system control model is passive theory. When a system's power input is bigger than its output, the system is stable. But this method does not consider the system's internal characteristics and structure. A model was designed by computing the input and output in order to eliminate the network transmission time delay. But this model changes the system's characteristics and increases the time delay [1–4].

The control method based on the evasion of network transmission time makes the whole system no longer influenced by time, keeping the original system's stability. From stable theorem based on the event, we know that the system stability needs two premises. First, the original robot assembly system itself is stable and controllable. Second, it is not related time that designed reference model *s*. This model function must be monotone increasing along with time *t;* it guarantees that controls and feedback use the same event reference. To guarantee that controls and feedback use the same event reference, reference model *s* must meet the requirement mentioned above. For instance, it may put the operation distance or the sequence number as the reference. In Ref. [1], $s \in [1,2,3,\ldots]$; this method is actually time-delay discretion. It can maintain the original system's stability. But because of the time delay, there are significant differences between the far-end operator and the local operator. For a fixed-length time delay, many say the track performance is much better; but in an actual network, the time delay is variable, sometimes even losing bag or overtime. At the moment, the difference between the master–slave manipulator position and the force track is quite large [5–8].

Figure 28.1 depicts a flowchart of an event-based robot teleoperation bidirectional force feedback control system. In this system, the master teleoperator had the force feedback function control handle in order to obtain the telepresence. The slave teleoperator is the manipulator. In order to make the master–slave manipulator track well, it adopted the speed control way. The two flowcharts in Fig. 28.1 stand for the far-end client server and the robot server. The event reference is the sequence number. Note that the robot's original control system in Fig. 28.1 is invariable; the robot server end has two pieces of feedback information: One is the force, and the other is the image information. In Fig. 28.1, the force and the image are fed back together to the operator to maintain the image and the bilateral uniformity. In fact, the standard picture frame frequency is 25/s; therefore, the feedback times of the robot image data must be greater than the force feedback times. Because of the large amount of data, one needs to consider more effective compression algorithms and transport protocols.

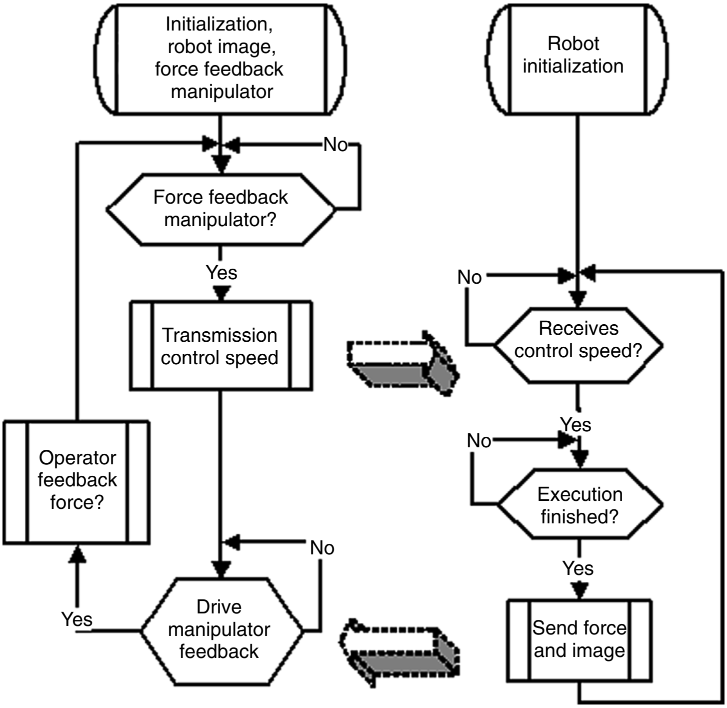

Initialization, robot image, force feedback manipulator
No
Force feedback manipulator?
Yes
Transmission control speed
Operator feedback force?
No
Yes
Drive manipulator feedback
Robot initialization
No
Receives control speed?
Yes
No
Execution finished?
Yes
Send force and image

Fig. 28.1. Flowchart of event-based teleoperation system with bilateral force feedback

28.2.1 Optimized control model

An event-based control model adopted a reference model that was not time-related and took it as the control basis. It shielded the time and guaranteed the system's stability, but the time-delay's influence still existed in the system. When system realized control way based on time, it must need to use the transmission-waiting the way. Each operation needs waiting time. Thus, the key idea of our improved algorithm is to reduce the waiting time.

28.2.1.1 Optimizing a control model based on an event

Based on the question mentioned above, we designed a new control model: a control network based on the event-designed event reference model *s*, which can slow down or smooth the time delay's influence. According to an examination network of the real-time state (QoS), the new control model joins one kind of forecast mechanism to enable the operator to respond quickly to the sensation operation force. After the actual force, position, and image are fed back, it simultaneously renovates the force and image. The system's design is shown in Fig. 28.2.

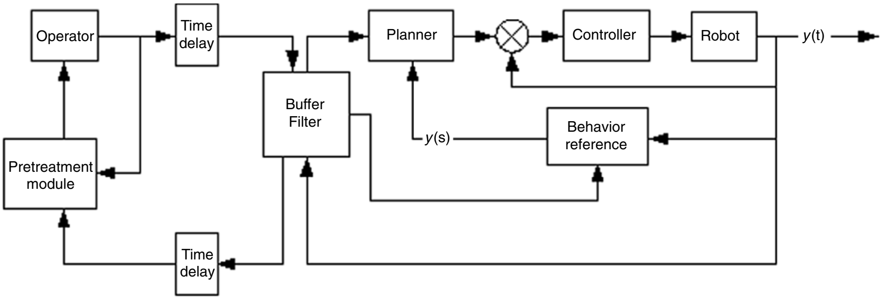


Fig. 28.2. An enhanced event-based teleoperation system

In the actual operation process, the remote operator needs to carry outthe decide the next movement according to its vision (image) and feeling (force feedback handle). Therefore, we considered vision when creating our design. Because the visual information is quite large, on the one hand, it can cause a serious time delay; on the other hand, it is very difficult to establish a model. The supervisor control mode based on an event is the same as in Ref. [2]. In an actual operation, the master control operator asks the system for permission to operate it. The operation is completely carried out according to the operator's request; the system not only responds fast, but also asks to be autonomous.

This model increased a simulation predictor at the operator's end. Its goal is to directly simulate the position or force, which must be operated according to the network time delay (including statistical network jam condition, bandwidth situation, and QoS). If the model received the actual feedback from the far-end robot, the actual data are adopted; otherwise, the simulation result is directly fed back to the operator. After the actual position or strength information is fed back in time, it is input to the predictor and processed. This function has two points: First, it revises the simulator and feed backs the real result and the simulation result given to the operator after a smooth handle; second, according to the returns result to order the sequence and the group information, it renews the transmission instruction queue. This process is the counterprocess of the far-end robot sequence processing module.

28.2.1.2 Predictor model

1. The position predictor is adopted in the following way:

$$p(t) = p^d(h(t)), \tag{28.1}$$

where $p^d(h(t))$ is the expectation position according to the time-delay model $h(t)$.

2. Computation force or force.

The computation force or the force in the mixture control is hard to compute. The position and the speed state predictor in Ref. [2] have been designed. In our model, the simulation force computation asked system performance is high. Because the system dynamic model is complex, the client-end procedure is limited to a kind of robot. Because of network reason, when the system does not respond, it can give the user a false strength, leading the user to believe that the system network does not have a problem. If the forecasting model is not ideal, then the influence on the system can be larger.

Regarding an unknown system, the method we used was to calculate a virtual force based on a simplifying model for the present system, for instance, assuming each arm to have a regular, calculable physique, and then making dynamic adjustments according to the robot operation force returned from the real environment and making equation $\widetilde{F} \le F_C$ less tenable. If the simulated computation force F_C is smaller than the actual force $\widetilde{F}$, it may guarantee forecasting the feedback force to gradually enhance the actual feedback strength.

28.2.2 Stability analysis

The previously stated prerequisite of an improved algorithm satisfied the control mode's stability principle based on an event. Namely, the original robot assembly system was stable:

$$\dot{x} = f(X(t), u(t), t). \tag{28.2}$$

When preparing the figures, it is very important that the technical requirements are taken into consideration in order to obtain high-quality print output.

$$\lim_{t\to\infty} e(t) = \lim_{t\to\infty}\left(X(t) - X^d\right) = 0 \tag{28.3}$$

In the improved model, an operator end (local) and a control end (remote) both have an order sequence. Both sequences are SL = (X_1L_d, X_2L_d, …, XL_d), SR = (X_1R_d, X_2R_d,…, XR_d), respectively. The reference sequences are $SL(i)$ = (1, 2, …, L), $SR(i)$ = (1, 2, …, R), respectively, where $L \geq R$. Suppose the time delay of the ith operation ($i \leq R$) is T_ci, the forecasting time delay is T_pi, but the smallest time interval that satisfies the stability condition is T_si. The event-based control method in Ref. [5] needs $T_si \leq T_ci$ to be satisfied. In improved models, if $T_pi \leq T_ci$, then when $T_si \leq T_pi$ is satisfied, the stability of the amended model will be obvious.

It is easily known that T_ci changes at random, but T_pi can be predicted according to the characteristic of the network time delay and QoS situation. We only guarantee that $T_si \leq T_pi,$ and we choose the shortest possible prediction time delay. The system place trace performance is fine and stabilization.

28.2.3 The improvement controls an event-based model procedure

In a real operation, the order's initial time and QoS execution time need to be recorded in each starting order. In this way, the robot execution end gets an important reference.

(1) On the telerobot end, on the basis of the original planner, the model increases its queue management and the filtration module. In fact, it is one kind of expansion of the planner. Regarding the remote operation robot, all far-end operations and all results feedback must pass through the sequence/filter module. The processing step is as follows:

Step 1: Wait to receive and deal with the order; determine whether or not there are orders in the sequence to be handled. If there are orders, go to step 2; otherwise, circulate. At this moment, the state is set up as vacant. After the order has been received and put into sequence, it must be numbered.

Step 2: Obtain the foremost effective order in the waiting sequence. Suppose the order number is *k*, and send out an operating order to the planner. Set the event state as the waiting order state at the same time.

Step 3: The robot deals with the corresponding order.

Step 4: The robot returns the handling results and order number *k* to the sequence.

Step 5: Based on feedback order *k* and the result, the processing module feeds back the robot image to the code stream and sends the processing result and operating order number in the group to the remote operator.

Step 6: The remote operator renews the image according to the feedback processing result, adjusts the position and feedback force, and carries out the following operation.

Step 7: Repeat step 1.

(2) On the distant operator's end, the processing flow is as follows:

Step 1: Through the operation's contact surface—the keyboard, mouse, or feedback force handle—the remote operator starts to operate the far-end robot; it sends out the position and the speed instruction to the forecast processing module queue, carries out the corresponding operation, and waits for the far-end robot's feedback.

Step 2: The forecast processing module puts the order into the queue according to the former network's statistical value and the current condition and forecasts the time-delay value of the instruction feedback. Under this time delay, ifno information was fed back, the simulated computation force was fed back to the teleoperator. At the same time, it informed the operator that the current force was the simulated computation force. The user can decide whether or not to operate slowly. If the actual processing result was fed back, first the forecast processing module needs to judge this instruction serial number and the group instruction number, the renewed instruction waiting queue, and the forecast model. Simultaneously, it compares the calculating force and returns the difference and value feedbacks to the remote operator.

Step 3: Repeat step 1

28.3 Simulation

We use a teleoperation control system with two degrees of freedom as an example.

The kinematics equation is

$$\begin{cases} x = l_1 c_1 + l_2 c_{12}, \\ y = l_1 s_1 + l_2 s_{12}, \end{cases} \tag{28.4}$$

where c_1 , s_1, c_{12}, and s_{12} stand for cos θ_1, sin θ_1, cos(θ_1+ θ_2), and sin(θ_1+ θ_2), respectively.

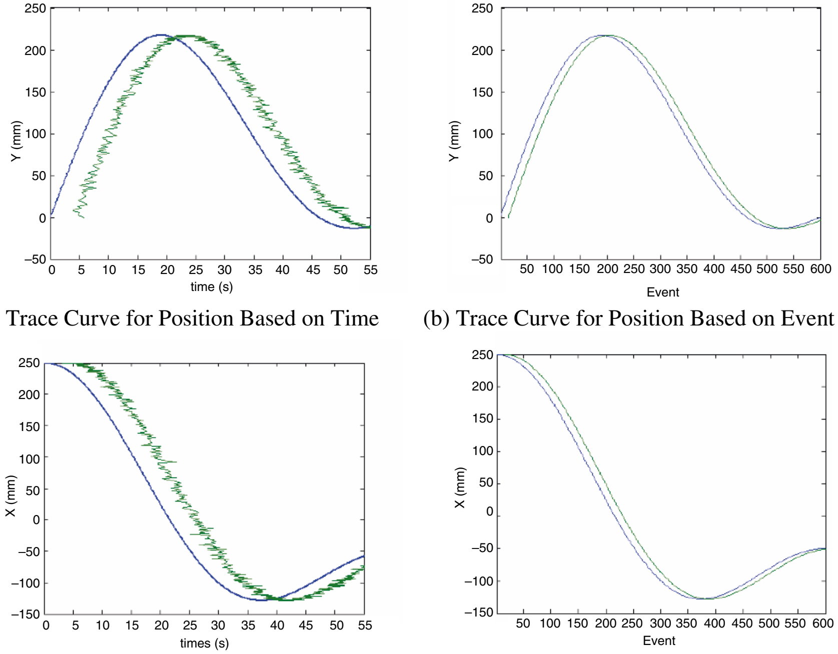
The kinematics equation is

$$\tau = D(q)\ddot{q} + h(q,\dot{q}) + G(q). \tag{28.5}$$

We adopt the PD control method,

$$\tau = D(q)J_h^{-1}(q)\left(V_1 - \dot{J}_h(q)\dot{q}\right) + h(q,\dot{q}) + G(q) + J_h^T(q)V_2. \tag{28.6}$$

Figure 28.3 shows the force curve comparison, based on the time and event positions.



(a) Trace Curve for Position Based on Time

(b) Trace Curve for Position Based on Event

(c) Trace Curve for Force Based on Time

(d) Trace Curve for Force Based on Event

Fig. 28.3. Simulation curves of 2-degree-of-freedom manipulator

A six-degree-of-freedom telerobot system simulation is adopted:

$$\tau = D(q)J_h^{-1}(q)\left(V_1 - \dot{J}_h(q)\dot{q}\right) + h(q,\dot{q}) + G(q) + J_h^T(q)V_2, \quad (28.7)$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} q \\ \dot{q} \end{pmatrix}, v = \ddot{q}^d + K_v(\dot{q}^d - \dot{q}) + K_p(q^d - q), \tau = \alpha(x) + \beta(x)v,$$

$$\alpha(x) = -D(q)J^{-1}(q)\left[J(\dot{q})\dot{q} - J(q)D^{-1}(q)\left(h(q,\dot{q}) + G(q)\right)\right], \quad (28.8)$$

$$\beta(x) = D(q)J^{-1}(q).$$

Based on the event control, the reference sequence is 1, 2, …. Controls the reference is

$$Y^d = \begin{pmatrix} y^d \\ \dot{Y}^d \\ \ddot{Y}^d \end{pmatrix} = \begin{pmatrix} s \\ v = \dfrac{ds}{dt} \\ a = \dfrac{dv}{dt} \end{pmatrix}. \quad (28.9)$$

Simulation mainly verifies the simulation force, the actual force (calculate with the PUMA560 library function), and the trace state of the master–slave hands under the time-delay forecast model. A simulation appears in Fig. 28.4 in detail. In order to contrast the measuring force curve of the forecast mechanism's control mode, we only list X - direction simulation force.

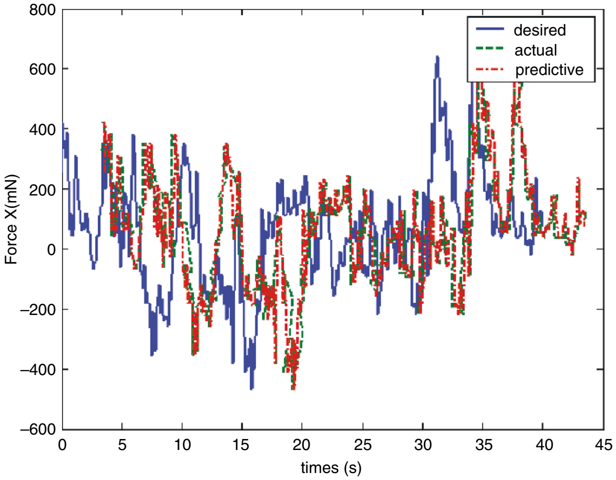


Fig. 28.4. Event-based curves of manipulation with predictive model

28.4 Conclusion

The time delay is a question that needs to be addressed and solved. This chapter first analyzed the existing control method teleoperation system. We proposed a buffer mechanism processing method on the basis of existing event-based telerobot systems. At the same time, we also proposed a design for the event-based reference state *s*. Finally, we carried out analysis and a simulation test of changed fixed-length time delay.

References

1. Imad HE (2002) Super media enhanced Internet-based realtime tele-robotic operations [PhD]. Michigan State University
2. Hu XD, Yu H, Chen Y (2002) Web-based data acquisition. J Zhejiang Univ Sci 2:135–139
3. Brady KJ (1997) Time-delayed control of telerobotic manipulators [DSc]. Washington University
4. Li X-M, et al. (2004) Hybrid event based control architecture for teler-obotic systems controlled through Internet. J Zhejiang Univ 5:296–302
5. Cobzas D, et al. (2003) Recent methods for image-based modeling and rendering. In: IEEE Virtual Reality 2003 tutorial 1, New York, pp 955–110
6. Akrivas G (2004) MPEG-4 Authoring tool using moving object segmentation and tracking. In: Video SHOT, Petros Daras, pp 239–246
7. Sichitiu ML (2001) Control of data networks: Models, stability and controllers [PhD]. University of Notre Dame
8. Ortega R, Chopra N, Spong MW (2003) A new passivity formulation for bilateral teleoperation with time delay. In: Advances in Time-Delay Systems, Workshop CNRS-NSF, Paris, pp 22–24

Chapter 29

Hardware implementation of a multidimensional signal analysis system

V.N. Ivanović, R. Stojanović, S. Jovanovski

Department of Electrical Engineering, University of Montenegro,
Cetinjski put bb., 81000 Podgorica, Montenegro,
very@cg.ac.yu, stox@cg.ac.yu, srdjaj@cg.ac.yu.

Abstract. Hardware implementation of the system for multidimensional (space/spatial–frequency (S/SF)) signal analysis is developed. Multiple clock cycle hardware implementation (MCI) of this system is proposed in [6]. Based on the two-dimensional S-method (2-D SM) definition and its relationship with the 2-D short-time Fourier transformation (STFT), the proposed system is developed. By sharing functional kernel, known as the STFT-to-SM gateway, [5,7], within the execution, developed design optimizes critical design performances of the multidimensional systems, such as hardware complexity, energy consumption, and cost. Moreover, it can implement almost all commonly used S/SF distributions (S/SFDs).

Keywords. Space/spatial–frequency signal analysis, Multiple clock cycle hardware implementation

29.1 Introduction

Conventional tools used in time (space/spatial)–frequency signal analysis, the spectrogram (SPEC) and the pseudo-Wigner distribution (WD), exhibit serious problems: low SPEC concentration around analyzed signals' instantaneous (local) frequency and the emphatic interference effects in the case of multicomponent signal analysis by using WD. These problems seriously

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_29,

limit applicability of these conventional tools. Consequently, almost all methods, proposed in the past two or three decades, are defined to retain high resolution of the WD and, at the same time, to alleviate interference effects when the multicomponent signals are analyzed. The SM [12,13] represents a very successful and very popular [3,5]–[7,9]–[11, 14–17] attempt in overcoming the above-noted problems. In the case of multidimensional (m-dimensional) signal analysis, the SM can be written in the following vector notation, [6,17]:

$$SM(\vec{n},\vec{k}) \\ =\sum_{\vec{i}} P(\vec{i})STFT(\vec{n},\vec{k}+\vec{i})STFT^{*}(\vec{n},\vec{k}-\vec{i}) \tag{29.1}$$

where $P(\vec{i})$ is the rectangular frequency domain (convolution) window, with $2L$+1 width in each directions, whereas $STFT(\vec{n},\vec{k})$ is the m-dimensional STFT of the analyzed m-dimensional signal $f(\vec{n})$, and $\vec{n}=(n_1,n_2,...,n_m)\in\mathbb{R}^m$.

Usage of the STFT, as an intermediate step in the SM definition, makes SM very attractive for implementation but, at the same time, quite numerical and time consuming. This significantly restricts its real-time applications. The hardware implementation, if possible, can overcome this nuisance. Additionally, the SM includes the STFT and the WD as its marginal cases, obtained for minimal and maximal convolution $P(\vec{i})$ window width, respectively. However, it produces better results than these conventional methods regarding some essential demands, such as calculation complexity, cross-terms reduction, and noise influence suppression [12,17,14,13].

For a long period of time, having in mind the technology limitations in the hardware design, only the 1-D systems for TF signal analysis are considered, usually in their single clock cycle (parallel) implementation (SCI) forms [8,9,15,16]. They are quite complex and require duplication of the basic calculation elements when they are employed more than once. In [5–7] the MCI hardware design that overcomes drawbacks of parallel architectures from [8,9,15,16] has been proposed.

Recently, the demands for development of the multidimensional systems have increased. Such systems are more complex than the 1-D ones and often could not be realized: the chip dimensions, power consumption, and cost are significantly increased, while the processing speed is lowered. In [6] we propose a way to extend the 1-D MCI architecture to the 2-D case.

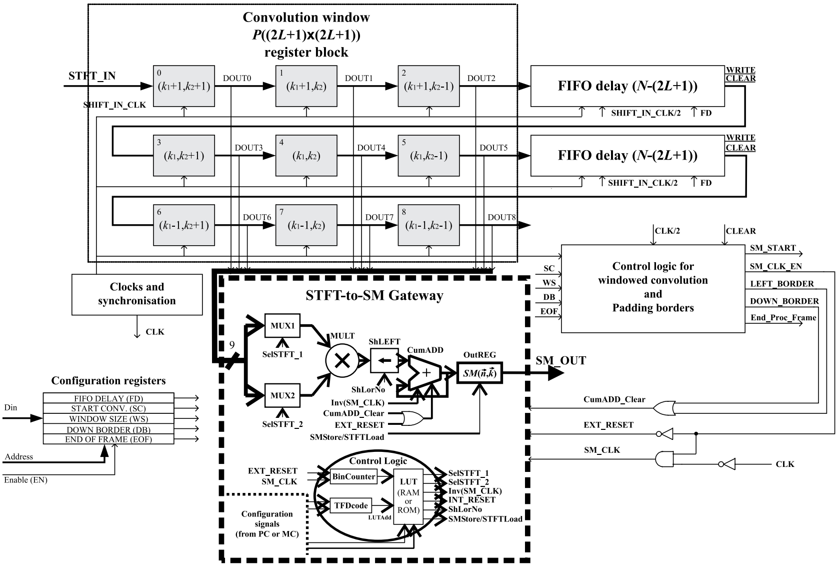


Fig. 29.1. Proposed MCI hardware design of the 2-D SM with L=1. In the center of the registers we denote the position of the stored 2-D STFT element in frequency plane, whereas the number in the left upper register's corner represents the address position of the corresponding 2-D STFT element at the STFT-to-SM gateway input multiplexers

The MCI architecture, proposed in [6], allows a functional kernel to be used more than once per S/SFDs execution, as long as it is used in different clock cycles. The abilities to allow S/SFDs to take different number of clock cycles and to share a functional kernel within the execution of a single S/SFD are highlighted as the major advantages of that design. Mentioned advantages optimize the hardware requirements. Using these possibilities, here, we realize S/SFDs using standard devices by developing the FPGA implementation of this system.

This chapter is organized as follows. After the introduction, the overview of the implemented architecture is presented. FPGA implementation of the 2-D system is developed in Section 29.3. In Section 29.4, the system implementation is tested and verified.

29.2 Overview of the implemented architecture

The system for S/SF signal analysis is based on the SM (29.1). Since the STFT is the complex transformation, the $SM(\vec{n},\vec{k})$ is calculated by independent processing of the $STFT(\vec{n},\vec{k})$ real and imaginary parts [5–7], [8,9,15]. Then, (29.1) involves only real multiplications and it is adapted for real-time hardware implementation. These parts of $SM(\vec{n},\vec{k})$ take the same form. In the 2-D domain case, this form is

$$\begin{aligned} SM_R(n_1,n_2,k_1,k_2) &= STFT_{\mathrm{Re}}^2(n_1,n_2,k_1,k_2) \\ &+2\sum_{i_1=0}^{L}\sum_{i_2=1}^{L} STFT_{\mathrm{Re}}(n_1,n_2,k_1+i_1,k_2+i_2) \\ &\quad \times STFT_{\mathrm{Re}}(n_1,n_2,k_1-i_1,k_2-i_2) \\ &+2\sum_{i_1=1}^{L}\sum_{i_2=0}^{L} STFT_{\mathrm{Re}}(n_1,n_2,k_1+i_1,k_2-i_2) \\ &\quad \times STFT_{\mathrm{Re}}(n_1,n_2,k_1-i_1,k_2+i_2) \end{aligned} \tag{29.2}$$

Equation (29.2) gives the 2-D SM for the point (k_1, k_2) of a 2-D frequency plane. It involves $CN(L)$ $=1+(L+1)L+L(L+1)=2L^2+2L+1$ summation terms (which will correspond to the number of clock cycles ($CN(L)$) in MCI), obtained by multiplying 2-D STFT elements that are symmetrically distributed around the (k_1, k_2) point in the 2-D frequency plane.

The 2-D SM hardware implementation, shown in Fig. 29.1, is done through its real computational line, since the imaginary one is identical. The design principle follows the developed form of (29.2), where each summation term is executed during the corresponding step (which takes one clock cycle). During the first clock, when L=0, the 2-D SPEC is executed from the 2-D STFT element, $STFT(\vec{n},\vec{k})$, situated in the middle point of the convolution window. Residual summation terms, for increased indexes i_1 and/or i_2, are obtained in the next steps (second, third, etc.). This improves the S/SFD concentration, aiming to achieve the one obtained by the 2-D WD. Note that the 2-D SM with arbitrary L requires $CN(L)$ clock cycles (by each point (k_1, k_2) of the 2-D frequency plane) to be executed. By breaking the S/SFDs execution into clock cycles, we are able to balance the amount of work done in each cycle, resulting in minimization of the clock cycle time. The presented hardware consists of two main parts: the convolution window register file and the STFT-to-SM gateway. The convolution window register file represents the hardware implementation of the 2-D convolution window function. It determines the order of the 2-D STFT input element addresses for which the corresponding 2-D SM output will be computed according to the algorithm (29.2). The STFT-to-SM gateway is used for hardware realization of this algorithm. It modifies the 2-D STFT elements obtained from the convolution window register file, in order to produce an improved concentration around local frequency based on the 2-D SM. STFT-to-SM gateway realizes 2-D SM calculation independently on the convolution window widths *L*, allowing the implemented S/SFDs to take different numbers of clock cycles for their calculation. This is made possible by sharing STFT-to-SM functional units for different inputs in different steps (clock cycles) that are controlled by the set of control signals (see Fig. 29.1. Details can be found in [7]). These abilities lead to minimization of the critical performances of the multidimensional systems: hardware complexity, energy consumption, and cost.

29.3 FPGA implementation approach

The module consisting of nine registers, noted as "convolution window register block," simulates window's sliding over the input 2-D STFT elements. Signal STFT_IN represents a 2-D STFT input data. *SHIFT_IN_CLK* clock signal enables registers' loading in appropriate period of time. Sliding of the window over the input signal for one position left is done by loading one STFT element STFT_IN per clock cycle *STFT_IN_CLK*. Then, each element of the convolution window register block row k_1+1 (as well as rows k_1, k_1-1) is shifted by PIPO (parallel-in-parallel-out) shift registers to generate data in time index (k_2, k_2-1). FIFO delay blocks are used to generate data of the convolution window register block column k_2+1 in time index (k_1, k_1-1). Note that the period of *STFT_IN_CLK* must be at least $CN(1) = 5$ times greater than the period of system clock *CLK*, in order to enable the corresponding SM calculation in 5 CLKs. Convolution operations inside the frame are managed by data arrangement part, which is called "control logic for windowed convolution and padding borders." The task of this block is to generate signals *SM_START*, *SM_CLK_EN*, *LEFT_BORDER*, *DOWN_BORDER*, and *End_Proc_Frame* considering input parameters derived from frame size *N* and window size *L*. These parameters are stored in the "configuration registers" module, Fig. 29.1, and their values are as follows: $FD{=}N{-}(2L{+}1)$, $SC{=}2LN{+}(2L{+}1){-}1$, $WS{=}2L{+}1$, $DB{=}(N{-}2L){\times}N$, $EOF{=}N{\times}N{-}1$, [6]. Considering the input parameters as well as synchronization conditions related to the main clock signals *CLK* and *SHIFT_IN_CLK*, the signals *SM_CLK_EN*, *LEFT_BORDER,* and *DOWN_BORDER* manage the operation of the STFT-to-SM gateway by generating its control signals *CumADD_Clear*, *EXT_RESET,* and *SM_CLK* clock signal. The signal *SM_START* implicitly participates in the STFT-to-SM gateway operation managing through participating in generation of the other mentioned signals (*SM_CLK_EN*,

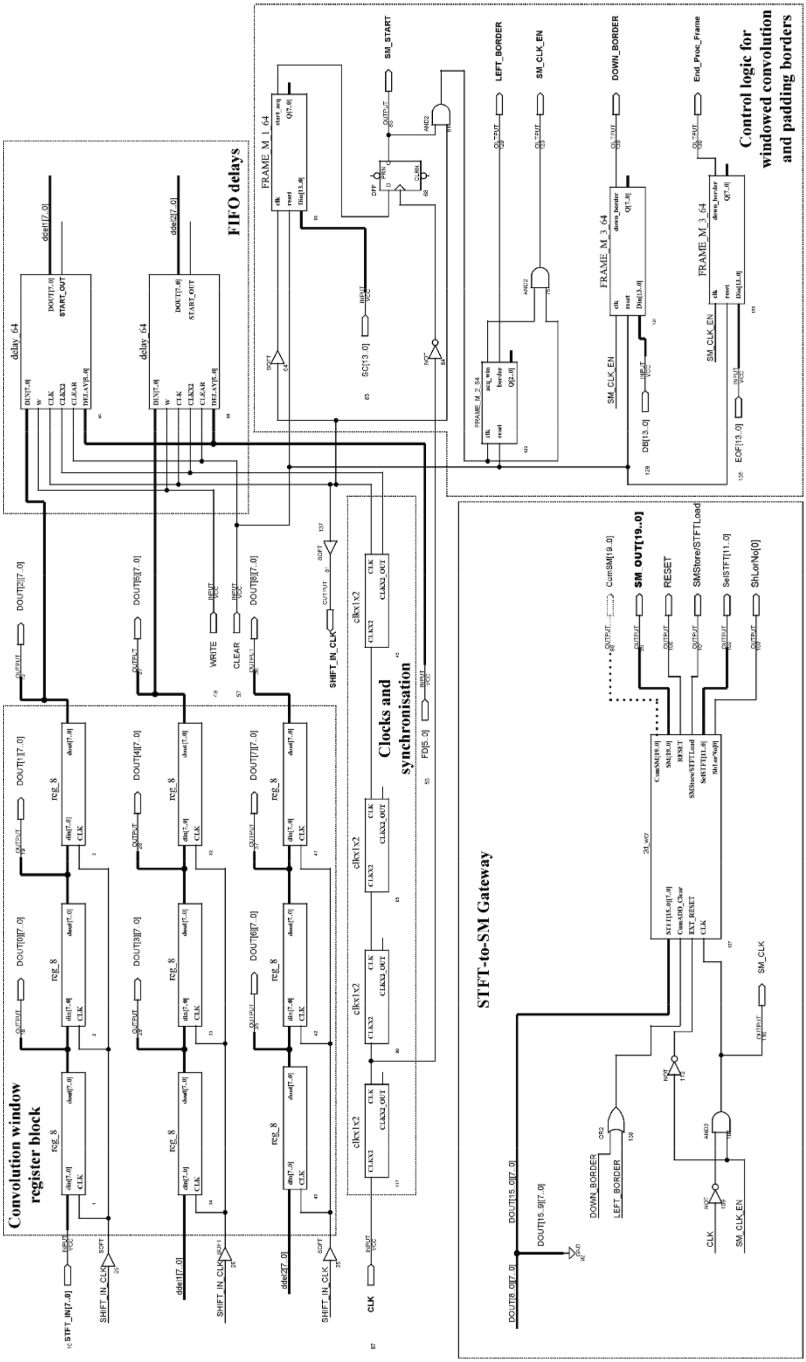


Fig. 29.2. Detailed schematic of developed FPGA implementation

LEFT_BORDER, *DOWN_BORDER*, *End_Proc_Frame*), whereas signal *End_Proc_Frame* indicates the end of the whole calculation process. Note that the *LEFT_BORDER* and/or *DOWD_BORDER* signals cause generation of the *CumADD_Clear* signal which resets cumulative adder integrated in STFT-to-SM gateway. For each window position, the *SM_CLK_EN* signal forwards the series of *SM_CLK*s that run SM calculation according to the algorithm given by Eq. (29.2). After $CN(1) = 5$ *SM_CLK*s, the SM will be calculated and stored in the output register. Additionally, the *SM_CLK_EN* signal resets the gateway when it takes zero value. When the window slides over the 2-D signal, the signals *LEFT_BORDER* and *DOWN _BORDER* are generated, to allow padding the borders of the frame with 0s.

In the FPGA implementation of the system, shown in Fig. 29.2, new library components were designed for “control logic for windowed convolution and padding borders” module. It consists of different FRAME_M_X modules for generating *SM_START*, *LEFT_BORDER*, *SM_CLK_EN*, *DOWN_BORDER,* and *End_Proc _Frame* signals, respectively. The basic components of these modules are variable length up–down counters with asynchronous reset and binary magnitude comparators. Each counter controls setting of the corresponding output signal from the “control logic for windowed convolution and padding borders” module by counting up to the appropriate parameter’s value from the “configuration registers.” The SM_START signal is generated to indicate that the system is ready for execution, considering input parameter SC (start convolution). Note that in the FPGA implementation, the *SM_CLK_EN* signal is used as clock signal for frames FRAME_M_3_64 that generate *DOWN_BORDER* and *End_Proc_Frame* signals.

The FPGA implementation scheme of the complete system is given in Figs. 29.2 and 29.3. The system units are implemented using the mixed approach allowed by design hierarchy, standard digital components from Altera’s libraries, AHDL-based mega functions, and developed VHDL-based modules. The adopted or developed VHDL or AHDL components have been parameterized in terms of input data size, horizontal and vertical depths of the FIFO delays, as well as of the window and image dimensions. Cascades of general latch registers (Altera’s 8dff) build convolution window register block. The FIFO delay is composed from Altera’s cycle-shared FIFO parameterized megafunction (CSFIFO) with added threshold – read request feedback.

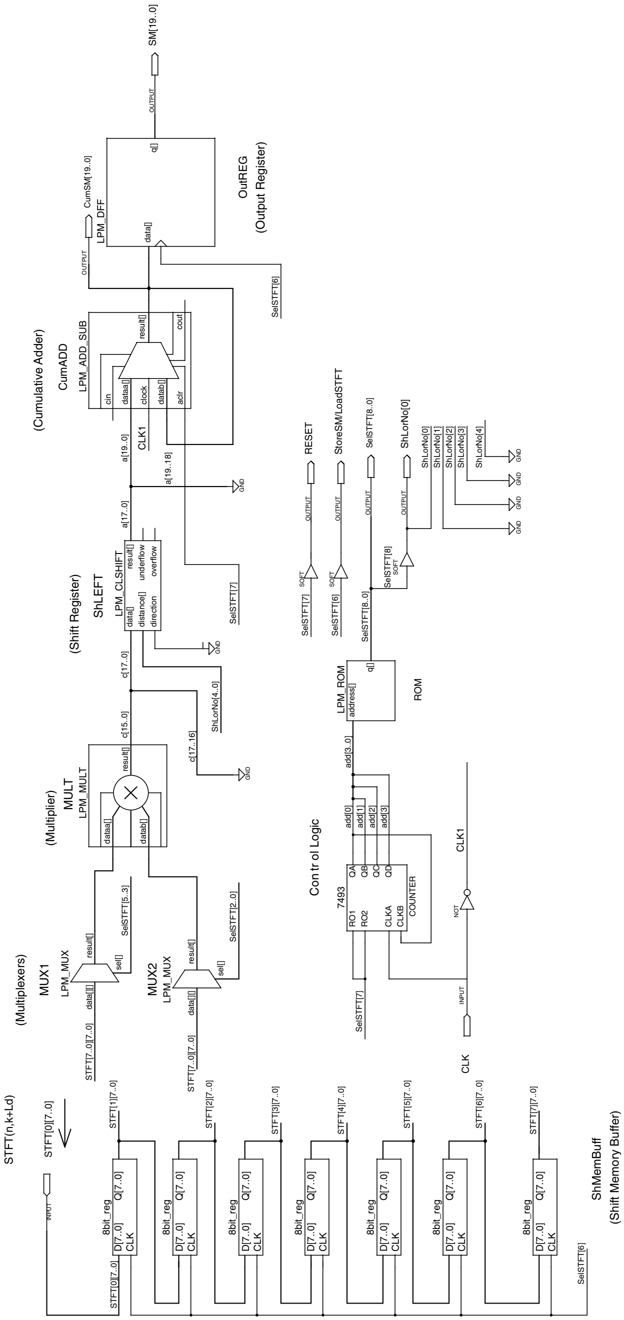


Fig. 29.3. The schematic of the 8-bit STFT-to-SM gateway implemented in FPGA

29.4 Testing and verification

In order to verify the chip operation, before its programming, the compilation and simulation have been performed by processing usually complex 2-D test signal:

$$f(x,y) = \cos[20\pi(x-0.75)^2 + 22\pi(y-0.75)^2] \\ + 0.5e^{j[-100\cos(\pi x/2)+100\cos(\pi y/2)]} \tag{29.3}$$

in the range $|x|< 0.75, |y|< 0.75,$ combined with the signal

$$f_s(x,y) = \cos\{1000\pi[(x+0.5)^2 + (y-0.5)^2]\} \tag{29.4}$$

whose, comparatively small, domain is $|x+y|< 0.1, |y-x-1|< 0.1$. We have applied the Hanning window in the 2-D STFT definition, whose widths along the x and y axes are W_x=W_y=1, respectively, and N=64. The computed 2-D STFT elements (their real and imaginary parts), normalized at the range [0, 255] and rounded to the 8-bit integers, are imported to the designed system input. Results of the real-time implementation are presented in Fig. 29.4, left-hand side. In order to verify the obtained results, numerical analysis, based on the same 2-D STFT elements, is performed and the results are presented in Fig. 29.4, right-hand side. Accuracy of the results obtained by using the designed system can easily be checked. Note that the results from Figure are computed at the point (x, y) = (−0.25,−0.25).

After simulation and verification, the Atlera's EPF10K20RC240-3 chip is configured by using the synthesized code [4]. It has 189 (51 input and 114 output) I/O pins. The rates of its 8-bit version silicon resources utilization are given in Table 29.1, first row. Additionally, Tables 29.1 and 29.2 give the comparison of two approaches (MCI and SCI) for different signal duration N×N (N=64 and N=256 are considered) and 3×3 convolution window. It can be easily noted that the occupation of the silicon resources, described by the total number of logic cells (LCs), is significantly less in the MCI case. The targeting devices are selected according to the optimal resource occupation. Consequently, used MCI devices have smaller capacity than the corresponding SCI ones. Naturally, the same device could be used for both approaches (MCI and SCI) provided that its maximal capacity is determined by the SCI requirements. Also, for both approaches (MCI and SCI), the LCs slightly vary with signal duration. On the other hand, the usage of memory bits significantly increases with signal duration. This is a consequence of the fact that delay functions are implemented by using FIFO memories. Precisely, for the N × N analyzed signal, the total

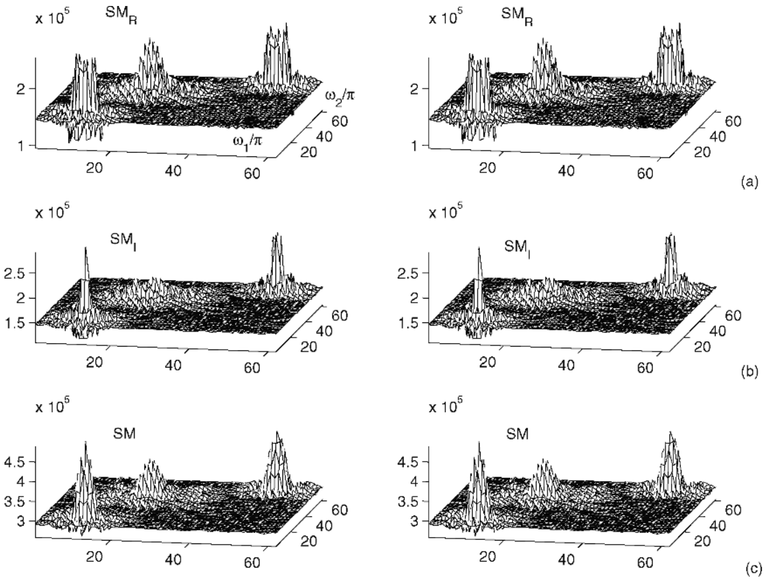


Fig. 29.4. The S/SF representation of the analyzed 2-D signal $f(x,y)+f_s(x,y)$ obtained by using the proposed hardware design (*left-hand side*), implemented in real FPGA devices (Altera's 10K series) and by numerical implementation (*right-hand side*)

Table 29.1. Utilized silicon resource (LCs, memory bits, utilized memory) for 8-bit 64×64 and 8-bit 256×256 2-D STFT-to-SM implementation

Design	Signal duration	Device	LCs	LCs utilized (%)	Memory bits	Memory utilized (%)
MCI	64×64	EPF10K20RC240-3	921	79	1216	9
SCI	64×64	EPF10K50RC240-3	2438	84	1024	5
MCI	256×256	EPF10K30BC356-3	965	55	4288	34
SCI	256×256	EPF10K50RC240-3	2478	86	4096	20

Table 29.2. Utilized silicon resource (embedded cells, EABs, required flip-flops) for 8-bit 64×64 and 8-bit 256×256 2-D STFT-to-SM implementation

Design	Signal duration	Embedded cells	Embedded cells utilized (%)	EABs	EABs utilized (%)	Flip-flops required
MCI	64×64	28	58	4	66	217
SCI	64×64	16	20	2	20	190
MCI	256×256	28	58	4	66	227
SCI	256×256	16	20	2	20	200

number of used memory bits would be expressed as $2 \times N \times 8 + MBS$, where *MBS* represents the number of memory bits used for the implementation of the Look-Up-Table from the STFT-to-SM gateway control logic, Fig. 29.1. Note that for SCI approach we have $MBS=0$.

29.5 Conclusion

FPGA implementation of the flexible system for S/SF signal analysis is presented. The system is based on the MCI of the 2-D SM. It allows the implemented S/SFDs to take different numbers of clock cycles and to share functional kernel, used to perform an S/SFD operation, within their execution. This property enables optimization of the critical design parameters.

Acknowledgment. This work is supported by the Ministry of Education and Science of the Republic of Montenegro.

References

1. Cohen L (1995) Time-frequency analysis. Prentice Hall, New Jersey
2. Dudgeon DE, Mersereau RM (1984) Multidimensional digital signal processing. Prentice Hall, New Jersey
3. Goncalves P, Baraniuk RG (1998) Pseudo affine Wigner distributions: Definition and kernel formulation. IEEE Trans. SP 6:1505–1517
4. Iborra A, Fernändez C, Älvarez B, Fernändez-Merono JM (2001) FPGA solution of low cost applications of real-time AVI systems. Dedicated Sys. to Mag. 2:793–84
5. Ivanović VN, Stanković LJ (2004) Multiple clock cycle real-time implementation of a system for time-frequency analysis. In: Proc. of the 12th EUSIPCO, Vienna, Austrija, pp 1633–1636
6. Ivanović VN, Stojanović R, Jovanovski S, Stanković LJ (2006) An architecture for real-time design of the system for multidimensional signal analysis. In: Proc. of the 14th EUSIPCO, Florence, Italy
7. Ivanović VN, Stojanović R, Stanković LJ (2006) Multiple clock cycle architecture for the VLSI design of a system for time-frequency analysis. EURASIP J. Appl. Signal Processing, Special Issue on Design Methods for DSP Systems, pp 1–18
8. Liu KJR (1995) Novel parallel architectures for Short-time Fourier transform IEEE Trans. CAS-II 12:786–789
9. Petranović D, Stanković S, Stanković LJ (1997) Special purpose hardware for time-frequency analysis. Electron. Lett. 6:464–466

10. Richard C (2002) Time-frequency-based detection using discrete-time discrete-frequency Wigner distribution. IEEE Trans. SP 9:2170–2176
11. Scharf LL, Friedlander B (2001) Toeplitz and Hankel kernels for estimating time-varying spectra of discrete-time random processes. IEEE Trans. SP 1: 179–189
12. Stanković LJ (1994) A method for time-frequency analysis. IEEE Trans. SP 1:225–22.
13 Stanković LJ, Ivanović VN, Petrović Z (1996) Unified approach to the noise analysis in the Wigner distribution and spectrogram. Ann. Telecomm. 11–12: 585–594
14 Stanković LJ, Stanković S, Djurović I (2000) Space/spatial-frequency analysis based filtering. IEEE Trans. SP 8:2343–2352
15 Stanković S, Stanković LJ (1997) An architecture for the realization of a system for time-frequency analysis. IEEE Trans. CAS-II 7:600–604
16 Stanković S, Stanković LJ, Ivanović VN, Stojanović R (2002) An architecture for the VLSI design of systems for time-frequency analysis and time-varying filtering. Ann. Telecomm. 9-10:974–995
17 Stanković S, Stanković LJ, Uskoković Z (1995) On the local frequency, group shift and cross-terms in the multidimensional time-frequency distributions; A method for multidimensional time-frequency analysis. IEEE Trans. SP 7:1719–1725

Chapter 30

Space-filling fractal microstrip antenna

M. Ismail, H. Elsadek, E. A. Abdallah, A. A. Ammar

Microstrip Department, Electronics Research Institute, El-tahrir St., Egypt

Abstract. The idea of space-filling properties of the fractal is used to minimize the area of a microstrip rectangular patch antenna operating at 2.45 GHz. The rectangular microstrip patch is used as an initiator to all iterations. The first four iterations of Koch at angles 30°, 60°, 80°, 90° and pulse 2.45 were studied. It is found that the second iteration of the proposed shape is the most suitable one because the third and fourth iterations have inferior antenna parameters. The size reduction in this case is about 15.1%. In order to further reduce the antenna size many trials have been carried out. These trials include adding a shorting wall, adding an air gap, adding a shorting wall and an air gap together, adding a shorting wall (or pins) together with an air gap and inverting the patch. As a consequence 46.12% reduction in size is obtained with reasonable antenna gain. The designed antennas were fabricated on teflon dielectric substrates and good agreement is found between simulated and measured results.

Keywords. Fractal, Microstrip antenna, Space-filling

30.1 Introduction

Fractals can be used to miniaturize patch elements as well as wire elements, due to their space-filling properties [1–5]. The same concept of increasing the electrical length of a radiator can be applied to a patch element. The patch antenna can be viewed as a microstrip transmission line [6–8]. Therefore, if the current can be forced to travel along the convoluted

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_30,

path of a fractal instead of a straight euclidean path, the area required to occupy the resonant transmission line can be reduced. This technique has been applied to patch antennas in various forms [9–11].

The rectangular microstrip patch is used as an initiator to all iterations. The first four iterations of Koch at angles 30°, 60°, 80°, 90° and pulse 2.45 were studied. In order to enhance the antenna bandwidth and other antenna parameters a modified pulse 2.45 microstrip patch antenna was suggested. The examined patch is 46.12% shorter than the simple rectangular patch. The designed antennas using the ready-made software package (Zeland IE3D) [12] were then fabricated using thin film technology and photolithographic technique and their input impedances and reflection coefficients were measured using a vector network analyzer in the required frequency range.

30.2 Initiator of Koch and pulse 2.45 antenna

A rectangular patch antenna was designed to resonate at Bluetooth frequency 2.45 GHz on dielectric substrate with $\varepsilon_r = 2.2$ (Duroid 5880) and h = 1.5748 mm. The antenna is fed by a probe coaxial feed at the position $x_0 = 0$ mm, $y_0 = 12.46$ mm from the bottom edge. Simulating the rectangle structure using Zeland IE3D to obtain the reflection coefficient ($|S_{11}|$ in dB) and the radiation pattern gives the results shown in Fig. 30.1(a,b,c). Tables 30.1 and 30.2 show the resonant frequency, −10 dB impedance bandwidth, and the performance parameters of the antenna.

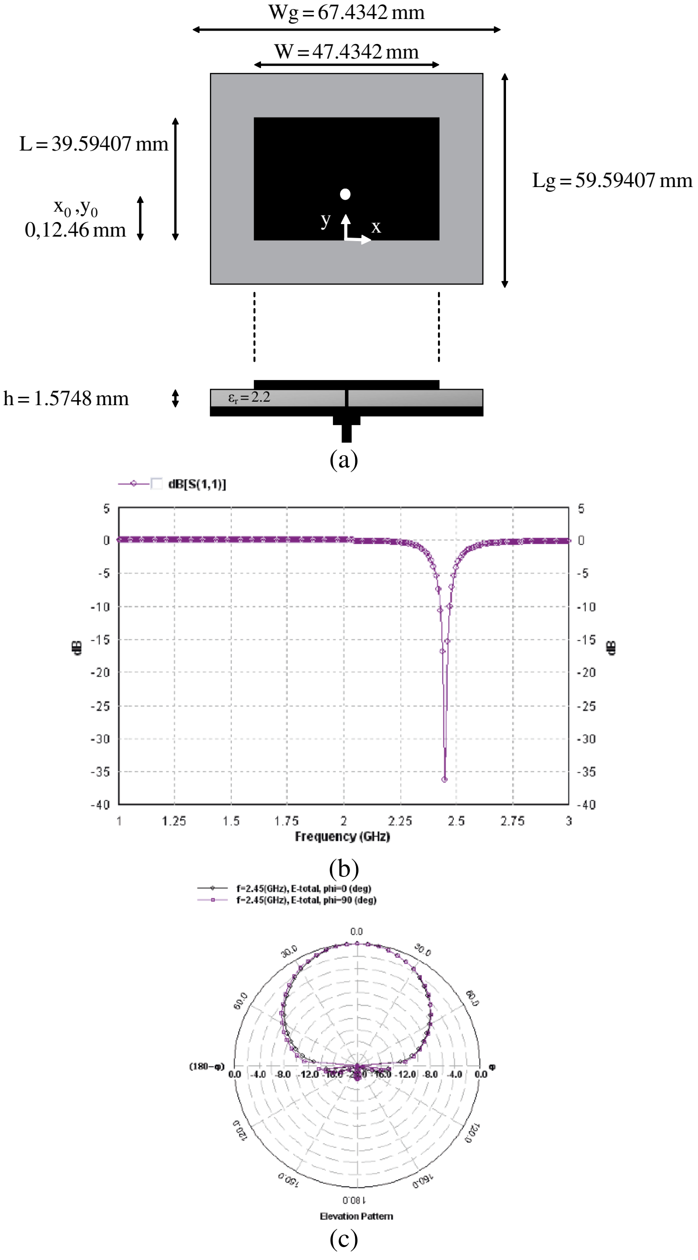
Table 30.1. Resonant frequency, reflection coefficient, and bandwidth for the initiator of pulse 2.45 antenna

F in GHz	$\mid S_{11} \mid$ in dB	BW in MHz	BW (%)
2.45	−36.15	44.1	1.8

Table 30.2. Antenna parameters for the initiator of pulse 2.45 antenna

Parameters	2.45 GHz
Gain (dBi)	7.12
Directivity (dBi)	7.6
Maximum (deg.)	(0,10)
3 dB beam width (deg.)	(79.9, 83.1)
Radiation efficiency (%)	89.03
Antenna efficiency (%)	89

From Tables 30.1 and 30.2, we can notice that the rectangle microstrip patch antenna may operate at the Bluetooth band, which has many applications. The rectangle microstrip patch antenna has narrow bandwidth and good radiation efficiency, gain, and directivity.



(a)

(b)

(c)

Fig. 30.1. (**a**) Initiator of pulse 2.45 antenna, (**b**) simulated $|S_{11}|$ in dB, and (**c**) simulated E- and H-plane radiation pattern

30.3 Koch microstrip patch antenna

Antennas in this section are based upon the fractal curve ‘Von Koch,’ a self-similar, iteratively reducing shape with an indentation angle θ=30°, 60°, 80°, and 90°. This fractal pattern is applied on one dimension of an euclidean-shaped antenna like a rectangular microstrip patch antenna reducing the antenna size by about 38%.

30.3.1 Koch antenna with angle 30°

We simulated the first four iterations of the Koch microstrip patch antenna with angle 30° as shown in Fig. 30.2(a,b,c,d) using Zeland IE3D simulator to obtain the reflection coefficient (| S_{11} | in dB) and the radiation pattern. These results are shown in Figs. 30.3 and 30.4(a,b,c,d). Tables 30.3 and 30.4 show the resonant frequencies,−10 dB impedance bandwidth, percentage size reduction, and the performance parameters of the antenna.

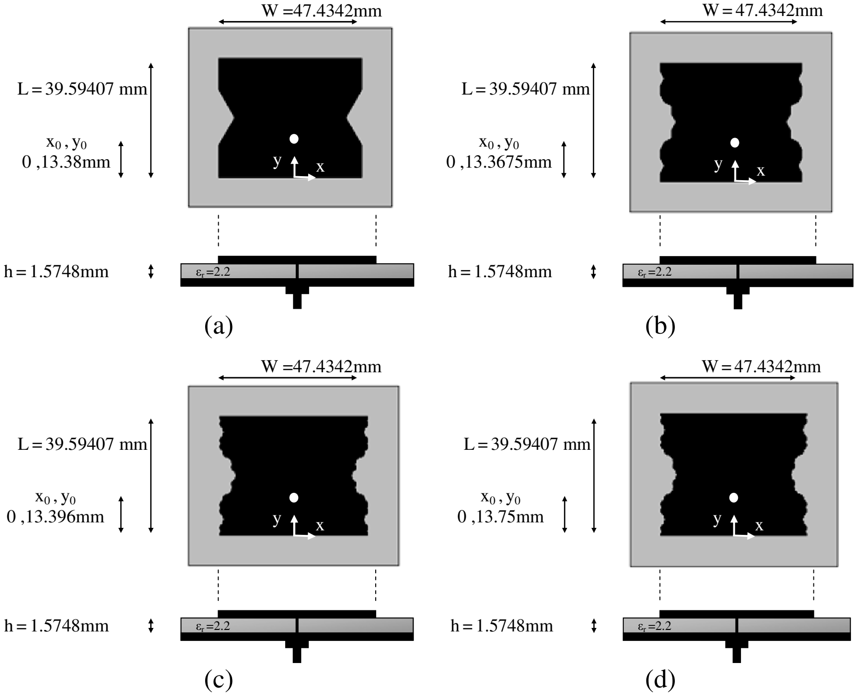


Fig. 30.2. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of Koch antenna with angle 30°

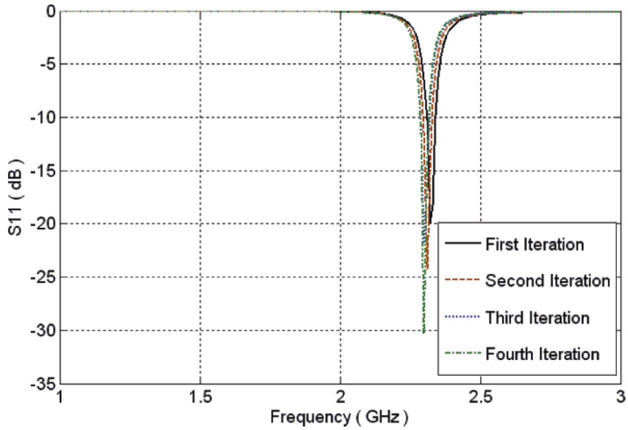
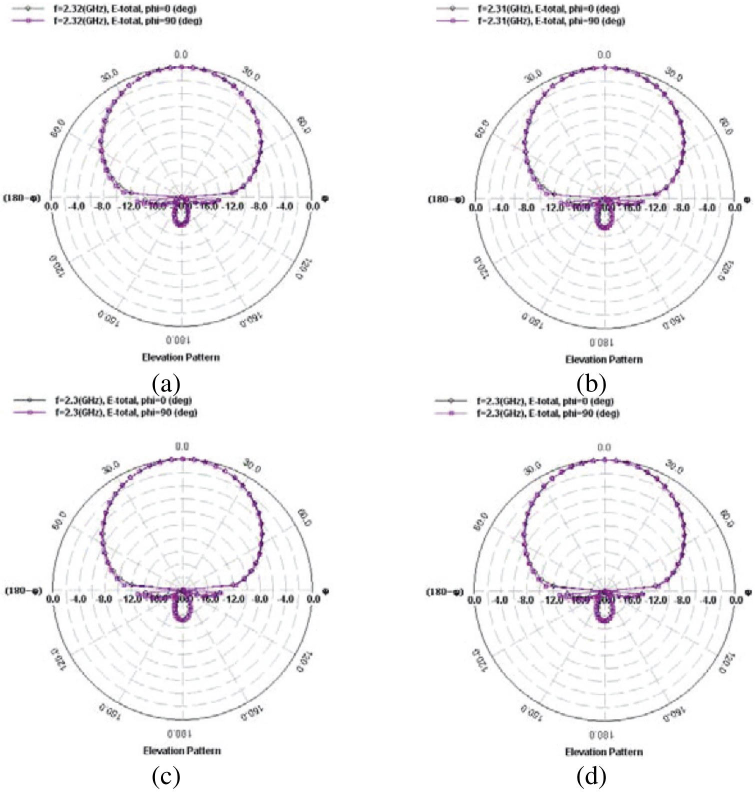


Fig. 30.3. Simulated | S_{11} | in dB for first, second, third, and fourth iterations of Koch antenna with angle 30°



(a) (b)

(c) (d)

Fig. 30.4. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of Koch antenna with angle 30° simulated E- and H-plane radiation patterns

Table 30.3. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the Koch antenna with angle 30° for the first four iterations

Iterations	F in GHz	$\vert S_{11} \vert$ in dB	BW in MHz	BW (%)	Size reduction (%)
1	2.32	−19.97	30.16	1.3	5.3
2	2.31	−24.2	29.799	1.29	5.7
3	2.3	−21.87	28.98	1.26	6.1
4	2.3	−30.37	28.06	1.22	6.1

Table 30.4. Antenna parameters of the Koch antenna with angle 30° for the first four iterations

Parameters	Frequency (GHz)			
	2.32	2.31	2.3	2.3
Gain (dBi)	6.9	6.86	6.84	6.86
Directivity (dBi)	7.45	7.44	7.42	7.42
Maximum (deg.)	(0, 10)	(0, 10)	(0, 20)	(0, 240)
3 dB beam width (deg.)	(82.9, 84.7)	(83.2, 84.9)	(83.46, 85.1)	(83.5, 85.1)
Radiation efficiency (%)	88.1	87.85	88.1	87.9
Antenna efficiency (%)	87.2	87.52	87.5	87.8

From Tables 30.3 and 30.4, we can notice that the four iterations have approximately the same resonant frequencies, bandwidth, gain, and radiation efficiency. The maximum reduction in size is 6.1%.

30.3.2 Koch antenna with angle 60°

We simulated the first four iterations of the Koch microstrip patch antenna with angle 60° shown in Fig. 30.5(a,b,c,d) using Zeland IE3D simulator to obtain the reflection coefficient (| S_{11} | in dB) and the radiation pattern. The results are shown in Figs. 30.6 and 30.7(a,b,c,d). Tables 30.5 and 30.6 show the resonant frequencies, –10 dB impedance bandwidth, percentage size reduction, and the other performance parameters of the antenna.

Table 30.5. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the Koch antenna with angle 60° for the first four iterations

Iterations	F in GHz	| S11 | in dB	BW in MHz	BW (%)	Size reduction (%)
1	2.05	−17.475	18.04	0.88	16.3
2	2.01	−37.37	17.085	0.85	18
3	1.98	−25.12	16.038	0.81	19.2
4	1.96	−32.5	15.876	0.81	20

Table 30.6. Antenna parameters of the Koch antenna with angle 60° for the first four iterations

Parameters	Frequency (GHz)			
	2.05	2.01	1.98	1.96
Gain (dBi)	6.24	6.16	6.05	5.99
Directivity (dBi)	7.14	7.08	7.04	7.01
Maximum (deg.)	(0, 60)	(0, 0)	(0, 150)	(0, 240)
3 dB beam width (deg.)	(86.3, 88.1)	(86.54, 88.59)	(86.8, 88.87)	(86.99, 89.04)
Radiation efficiency (%)	82.8	80.96	79.9	79.12
Antenna efficiency (%)	81.3	80.95	79.7	79.07

From Tables 30.5 and 30.6, we can notice that the four iterations have approximately the same narrow bandwidth, the same gain, and radiation efficiency. The maximum size reduction is 20% obtained in the case of the fourth iteration. The back radiation increases when the angle θ changes from 30° to 60°.

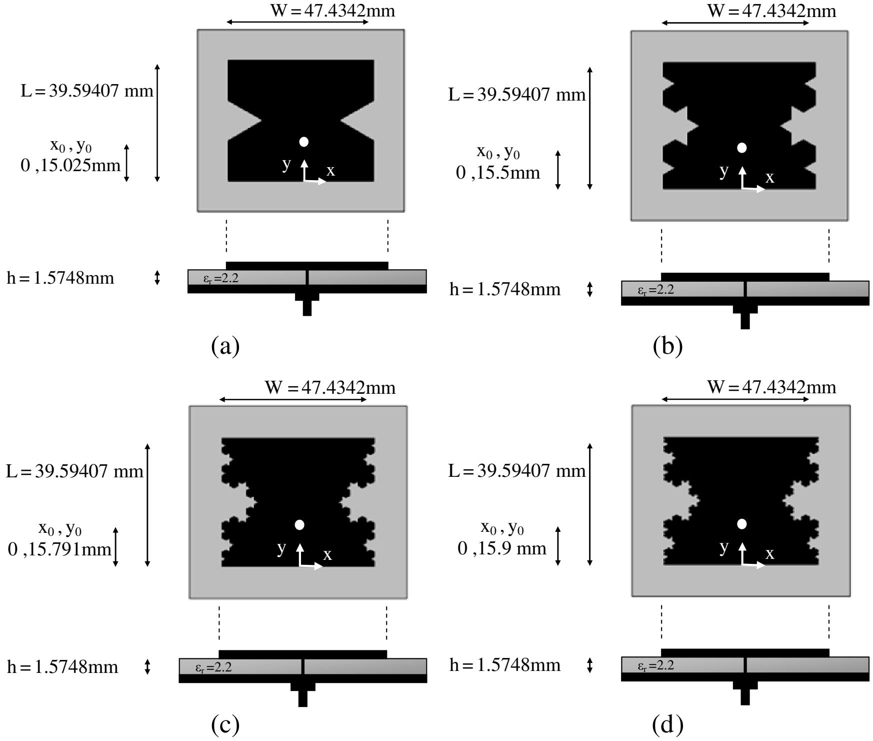


Fig. 30.5. (**a**) First iteration, (**b**) secnd iteration, (**c**) third iteration, and (**d**) fourth iteration of Koch antenna with angle 60°

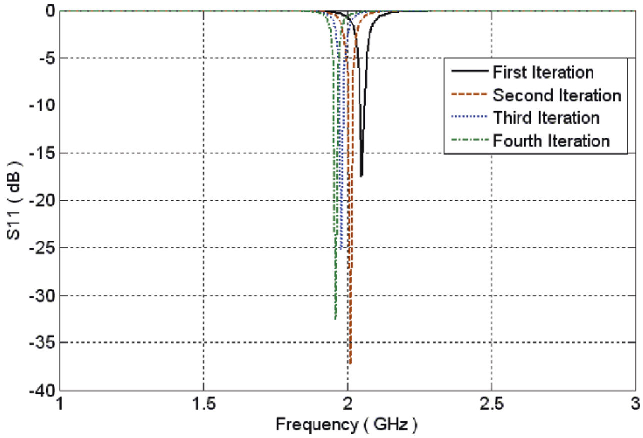
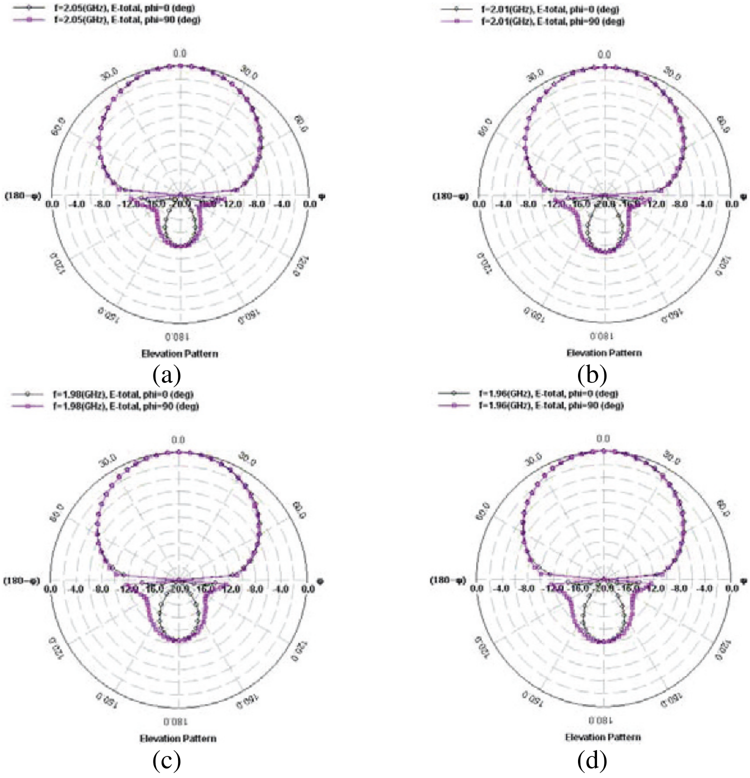


Fig. 30.6. Simulated | S_{11} | in dB for first, second, third, and fourth iterations of Koch antenna with angle 60°



(a)

(b)

(c)

(d)

Fig. 30.7. (a) Firth iteration, **(b)** second iteration, **(c)** third iteration, and **(d)** fourth iteration of Koch antenna with angle 60° simulated E- and H-plane radiation patterns

30.3.3 Koch antenna with angle 80°

We simulated the first four iterations of the Koch microstrip patch antenna with angle 80° shown in Fig. 30.8(a,b,c,d) using Zeland IE3D simulator to obtain the reflection coefficient (| S_{11} | in dB), the radiation pattern, and the antenna parameters. The results are shown in Figs. 30.9, and 30.10 (a,b,c,d) and Tables 30.7 and 30.8.

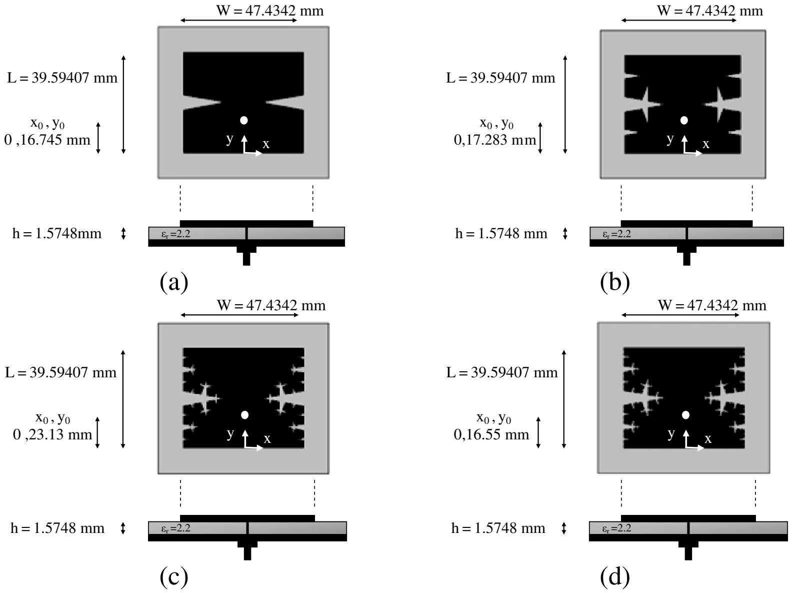


Fig. 30.8. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of Koch antenna with angle 80°

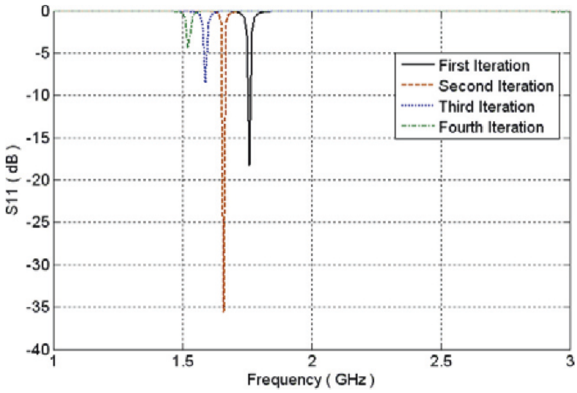
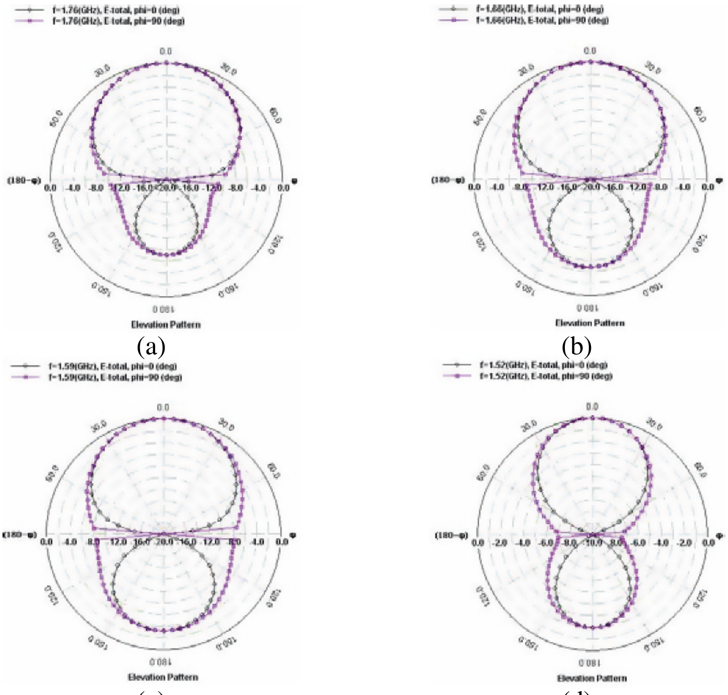


Fig. 30.9. Simulated | S_{11} | in dB for first, second, third, and fourth iterations of Koch antenna with angle 80°

Table 30.7. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the Koch antenna with angle 80° for the first four iterations

Iterations	F in GHz	$\vert S_{11}\vert$ in dB	BW in MHz	BW (%)	Size reduction (%)
1	1.76	–18.239	15.84	0.9	28.2
2	1.66	–35.6	9.462	0.57	32.2
3	1.59	–8.4961	–	–	35.1
4	1.52	–4.4183	–	–	38



(a) (b)

(c) (d)

Fig. 30.10. **(a)** First iteration, **(b)** second iteration, **(c)** third iteration, and **(d)** fourth iteration of Koch antenna with angle 80° simulated E- and H-plane radiation patterns

Table 30.8. Antenna parameters of the Koch antenna with angle 80° for the first four iterations

Parameters	Frequency (GHz)			
	1.76	1.66	1.59	1.52
Gain (dBi)	4.65	3.47	1.48	−1.14
Directivity (dBi)	6.6	6.2	5.7	5.17
Maximum (deg.)	(0, 20)	(0, 220)	(0, 340)	(0,10)
3 dB beam width (deg.)	(88.88, 89.8)	(88.42, 90.56)	(87.22, 92.17)	(86.26, 94.68)
Radiation efficiency (%)	64.23	53.3	43.59	36.65
Antenna efficiency (%)	81.3	80.95	79.7	79.07

From Tables 30.7 and 30.8, we can notice that the second, third and fourth iterations have poor gain and radiation efficiency. The third and fourth iterations have poor impedance matching. The maximum reduction size is 38%, but at the expense of other antenna parameters, namely bandwidth, gain, and radiation efficiency. The back radiation is very large.

30.3.4 Koch antenna with angle 90°

The Koch 90° was formed by dividing the side by 3 and moving to inside the patch about one-third the length. The probe feeding technique was used with different positions in each iteration. We simulated the first four iterations of the Koch microstrip patch antenna with angle 90° as shown in Fig. 30.11(a,b,c,d) using Zeland IE3D simulator to obtain the reflection coefficient (| S_{11} | in dB) and the radiation pattern. The results are shown in Figs. 30.12 and 30.13(a,b,c,d). Tables 30.9 and 30.10 show the antenna parameters.

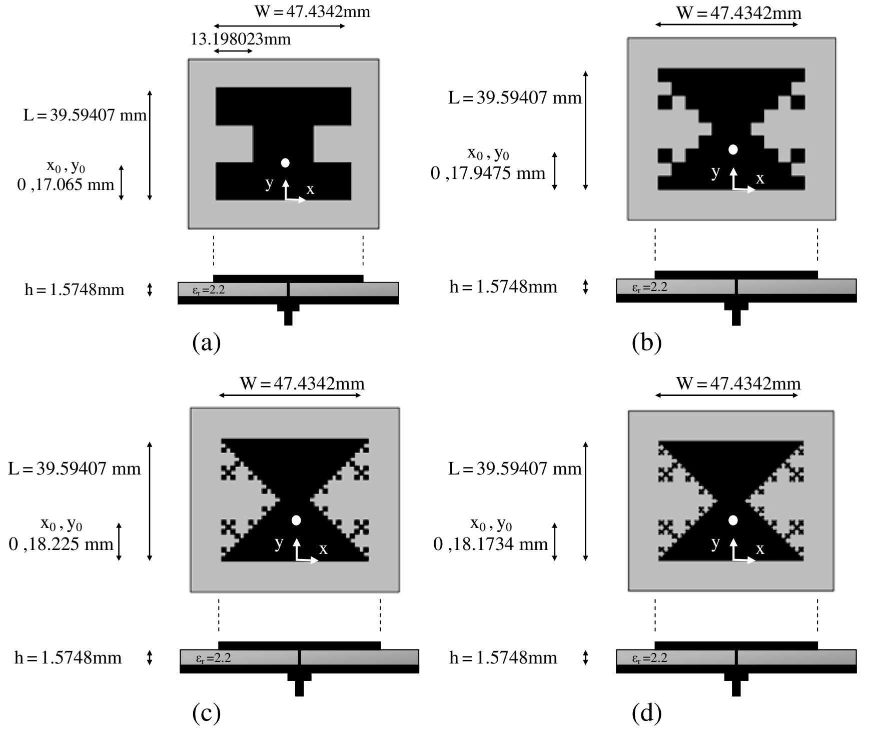


Fig. 30.11. (a) First iteration, **(b)** second iteration, **(c)** third iteration, and **(d)** fourth iteration of Koch antenna with angle 90°

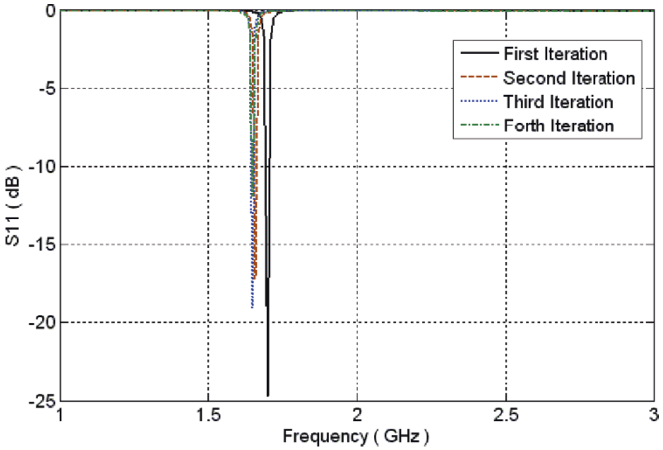
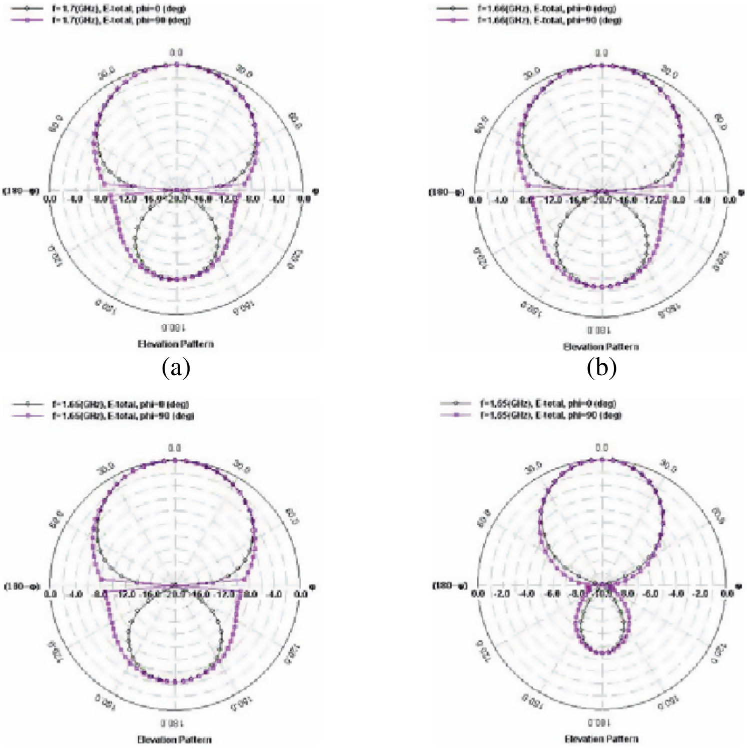


Fig. 30.12. Simulated | S_{11} | in dB for First, second, third, and fourth iterations of Koch antenna with angle 90°



(a)

(b)

(c)

(d)

Fig. 30.13. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of Koch antenna with angle 90° simulated E- and H-plane radiation patterns

Table 30.9. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the Koch antenna with angle 90°

Iterations	F in GHz	$\lvert S_{11} \rvert$ in dB	BW in MHz	BW (%)	Size reduction (%)
1	1.7	−24.745	12.75	0.75	30.6
2	1.66	−17.183	9.96	0.6	32.2
3	1.65	−19.0835	9.075	0.55	32.7
4	1.65	−11.893	3.63	0.22	32.7

Table 30.10. Antenna parameters of the Koch antenna with angle 90°

Parameters	Frequency (GHz)			
	1.7	1.66	1.65	1.65
Gain (dBi)	4.07	2.49	3.14	−4.57
Directivity (dBi)	6.38	6.19	6.11	6.07
Maximum (deg.)	(0, 280)	(0, 140)	(0, 70)	(0, 280)
3 dB beam width (deg.)	(88.86, 89.94)	(88.53, 90.49)	(88.38, 90.7)	(88.69, 90.79)
Radiation efficiency (%)	58.99	43.55	51.07	49.96
Antenna efficiency (%)	58.8	42.7	50.44	8.65

From Tables 30.9 and 30.10, we can notice that the four iterations have approximately the same resonant frequencies. The bandwidth reduces by increasing the order of iteration. The maximum reduction size is 32.7%, but at the expense of antenna gain and antenna efficiency which are very bad. The fourth iteration has narrower bandwidth, larger reflection, and very low efficiency. This iteration should be ignored. The back radiation is large and comparable to front radiation.

30.4 Pulse 2.45 antenna iterations

The first four iterations of the pulse 2.45 microstrip patch antenna are shown in Fig. 30.14(a,b,c,d). Zeland IE3D simulator was used to obtain the reflection coefficient (| S_{11} | in dB) and the radiation pattern, shown in Figs. 30.15 and 30.16(a,b,c,d). Tables 30.11 and 30.12 show the antenna figure-of-merits.

Table 30.11. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the pulse 2.45 antenna

Iterations	F in GHz	$\lvert S_{11} \rvert$ in dB	BW in MHz	BW (%)	Size reduction (%)
1	2.25	−18.32	27	1.2	8.2
2	2.08	−23.61	18.096	0.87	15.1
3	1.83	−19.85	15.738	0.86	25.3
4	1.51	−29.49	12.382	0.82	38.4

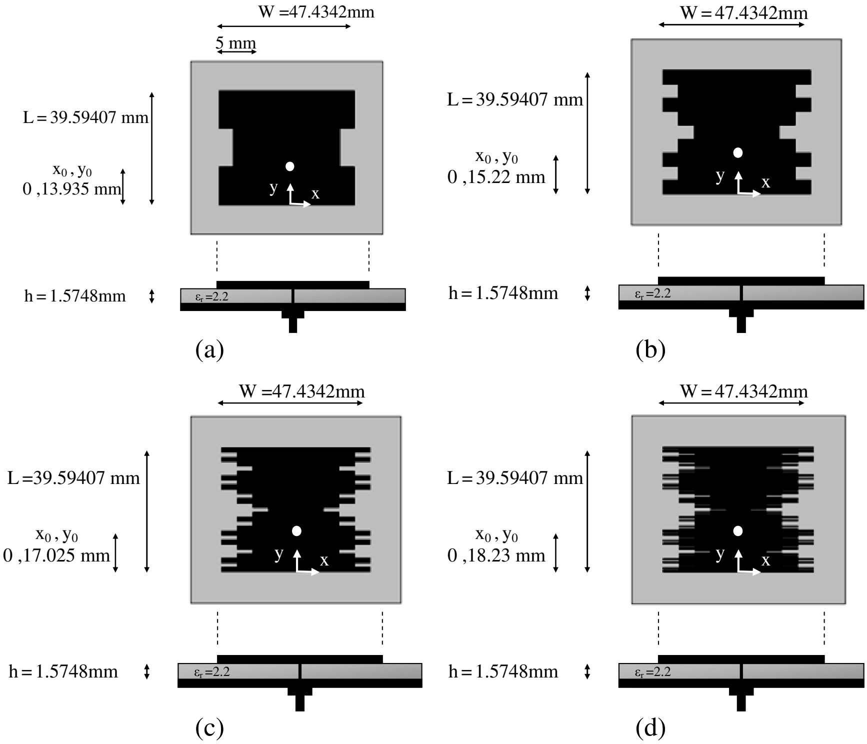


Fig. 30.14. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of pulse 2.45 antenna

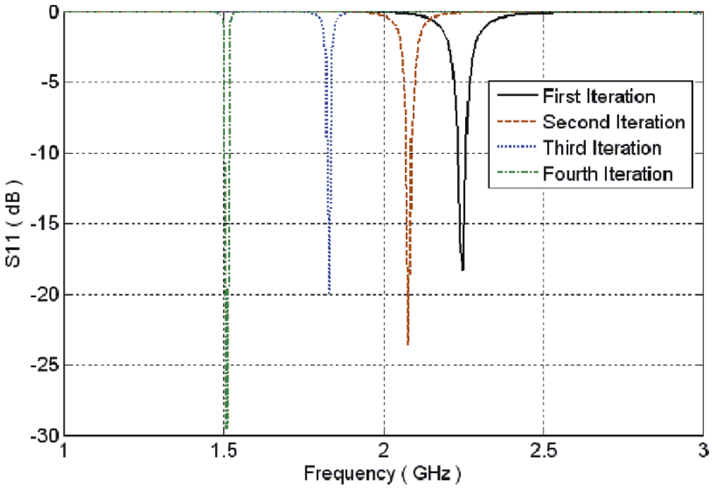


Fig. 30.15. Simulated | S_{11} | in dB for the first four iterations of pulse 2.45 antenna

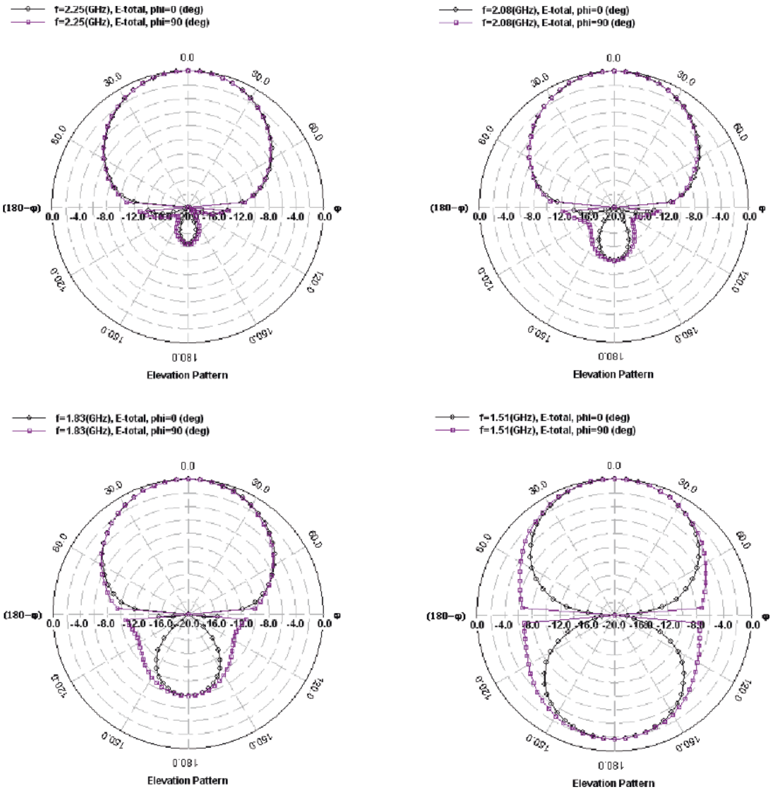


Fig. 30.16. (**a**) First iteration, (**b**) second iteration, (**c**) third iteration, and (**d**) fourth iteration of pulse 2.45 antenna simulated E- and H-plane radiation patterns

Table 30.12. Antenna parameters for the pulse 2.45 antenna

Parameters	Frequency (GHz)			
	2.25	2.08	1.83	1.51
Gain (dBi)	6.69	6.34	5.09	0.56
Directivity (dBi)	7.37	7.16	6.76	5.12
Maximum (deg.)	(0, 240)	(0, 110)	(0, 20)	(0, 350)
3 dB beam width (deg.)	(84.4, 85.55)	(85.99, 87.93)	(88.19, 89.69)	(86.18, 94.67)
Radiation efficiency (%)	86.76	83.04	68.8	35.03
Antenna efficiency (%)	85.48	82.68	68.12	34.99

From Tables 30.11 and 30.12, we can notice that the fourth iteration has very poor gain and radiation efficiency. The second, third, and fourth iterations have approximately the same bandwidth (very narrow bandwidth). The maximum reduction size is 38.4%, but at the expense of antenna parameters. The back radiation is very large comparable to the front one and increases with the iteration order. A bow-tie antenna with the same dimensions as that of the second iteration of pulse 2.45 antenna was simulated and the results of the two antennas were found to be very close.

30.5 The second iteration of pulse 2.45 antenna modification

We modified the second iteration of the pulse 2.45 microstrip patch antenna by using a shorting wall. The simulated | S_{11} | and radiation patterns without and with shorting wall are shown in Figs. 30.17(a,b), 30.18, and 30.19(a,b). Tables 30.13 and 30.14 show the resonant frequencies, −10 dB impedance bandwidth, percentage size reduction, and the performance parameters of the antenna namely gain, directivity, half-power beamwidth, radiation efficiency, and antenna efficiency.

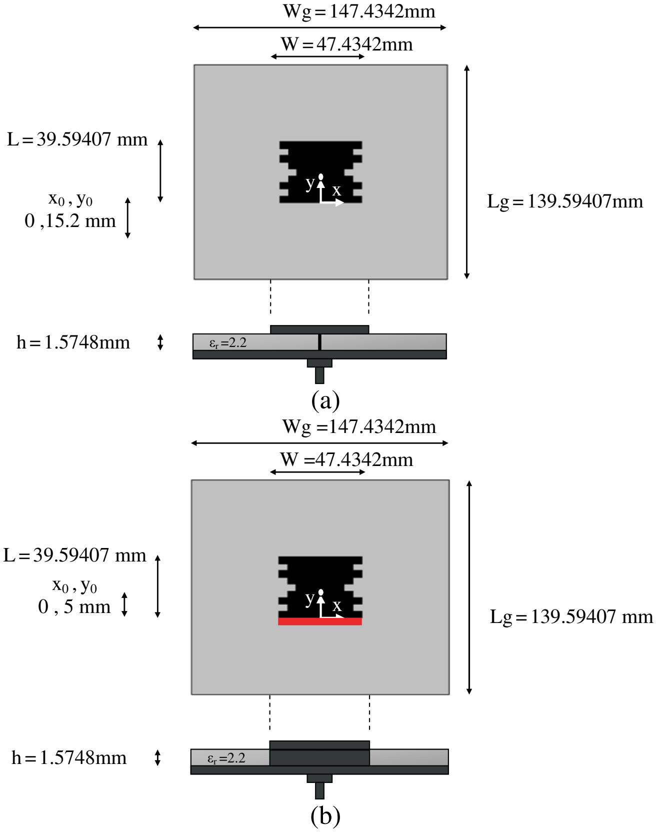


Fig. 30.17. (**a**) Second iteration without shorting wall and (**b**) with shorting wall

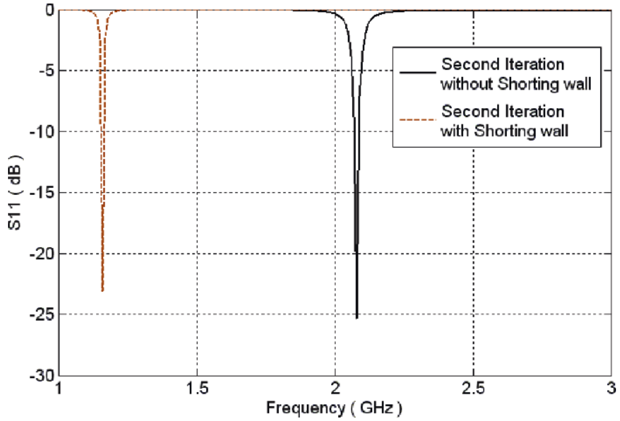
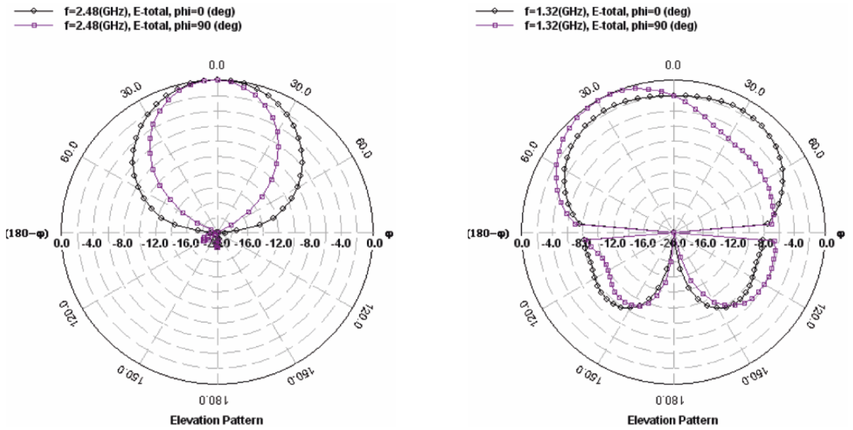


Fig. 30.18. Simulated | S_{11} | in dB for second iteration without and with shorting wall of pulse 2.45 antenna

Table 30.13. Resonant frequencies, reflection coefficient, bandwidth, and size reduction of the second iteration of the pulse 2.45 microstrip patch antenna without and with shorting wall

Second iteration	F in GHz	| S11 | in dB	BW in MHz	BW (%)	Size reduction (%)
Without shorting wall	2.48	−16.9	133	5.4	−1.2 (increased by 1.2)
With shorting wall	1.32	−34.8	66	5	46.12



(a) (b)

Fig. 30.19. (**a**) Second iteration without shorting wall and (**b**) with shorting wall of pulse 2.45 antenna with air gap simulated E- and H-plane radiation patterns

Table 30.14. Antenna parameters for the second iteration of the pulse 2.45 microstrip patch antenna without and with shorting wall

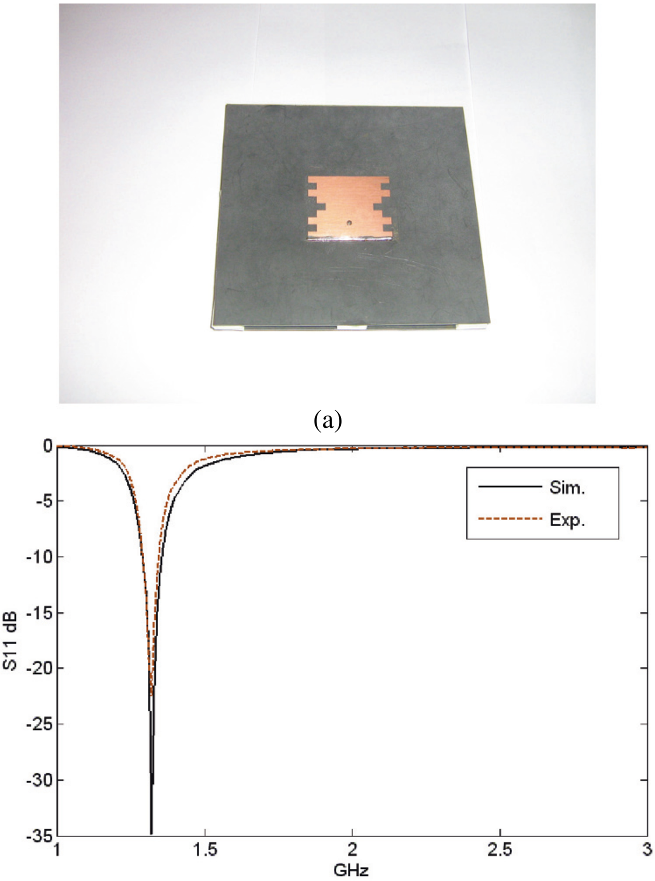
Parameters	Frequency (GHz)	
	2.48	1.32
Gain (dBi)	9.2	4.85
Directivity (dBi)	9.7	5.45
Maximum (deg.)	(0, 160)	(30 , 270)
3 dB beam width (deg.)	(51.35, 70.78)	(61.91 , 107.13)
Radiation efficiency (%)	91.45	87.12
Antenna efficiency (%)	89.6	87.1

From Tables 30.13 and 30.14, we can notice that the shorting wall gives reduction in size of approximately 46.12%. The directivity is reduced in the case of shorting wall as compared to the case without shorting wall, which is the reason for decreasing gain. The radiation pattern is distorted and becomes asymmetric due to the existence of the shorting wall at the antenna edge.

30.6 Experimental results

The second iteration with shorting wall of pulse 2.45 antenna with an air gap = 6.4 mm as shown in Fig. 30.20(a) is fabricated on a dielectric substrate covered with copper clad from both sides. The thickness of the copper layer is 35 μm. The dielectric substrate is RT/Duroid 5880, with relative permittivity $\varepsilon_r = 2.2$, dielectric height = 0.062 in. (1.5748 mm), and loss tangent tan $\delta = 0.0019$. The antenna performance was measured using Agilent 8719ES (50 MHz–13.5 GHz) vector network analyzer and was simulated using electromagnetic field solver IE3D (ZELAND) which adopts the method of moments. The computed and measured results were found to be in good agreement as shown in Fig. 30.20(b) and Table 30.15.

As shown in Table 30.15, the measurement and simulation results give good agreement with average normalized error equal to 0.02% in calculating F, and the size reduction is 46.12% as compared to the initiator. The measured reactive part of the input impedance of the antenna (capacitive due to the air gap) is larger than that simulated, while the radiation resistance is lower. The simulated value of the reflection coefficient is much better than the measured value due to many factors which were not taken into account.



(a)

(b)

Fig. 30.20. (**a**) Fabricated second iteration with shorting wall of pulse 2.45 antenna with an air gap = 6.4 mm. (**b**) Comparison between the simulated and measured $|S_{11}|$

Table 30.15. Resonant frequencies and BWs of the second iteration with shorting wall of pulse 2.45 antenna with air gap = 6.4 mm

		Simulated results		
fn (GHz)	| S11 | (dB)	BW(%)	Zin (Ω)	
			Real	Imaginary
1.32	−34.858	5	51.77	−0.5
		Experimental results		
fn (GHz)	| S11 | (dB)	BW(%)	Zin (Ω)	
			Real	Imaginary
1.3198	−22.25	4.09	43.55	−2.7

30.7 Conclusion

This chapter described the space-filling property of the fractal microstrip patch antenna. The iterations give maximum reduction in size equal to 46.12%. The fundamental limitation in fabricating the antenna is given by the resolution of the photoetching process. The fundamental resonant frequency decreases when the number of iterations increases. The difference between the resonant frequencies of the third and fourth iterations is so small.

References

1. Borja C, Romeu J (2003) On the behavior of Koch island fractal boundary microstrip patch antenna. IEEE Trans. Antennas Propagation 51:2564–2570
2. Gianvittorio J, Samii YR (2002) Fractal antennas: a novel antenna miniaturization technique, and applications. IEEE Antennas and Propagation Magazine 44:20–36
3. Guterman J, Moreira AA, Peixeiro C (2007) Dual-band miniaturized microstrip fractal antenna for a small GSM1800 + UMTS mobile handset. Proceeding of 12th IEEE Mediterranean Electrotechnical Conference, pp 499–501
4. Tedjini S, Vuong TP, Beroulle V (2005) Antennas for RFID tags. Proceeding of Smart Objects and Ambient Intelligence Conference 121:19–22
5. Werner DH, Gangul S (2003) An overview of fractal antenna engineering research. IEEE Antennas and Propagation Magazine 45:38–57
6. Cohen N (1997) Fractal antenna applications in wireless telecommunications. In: Proceedings of Electronics Industries Forum of New England, pp 43–49
7. Rahim MKA, Aziz MZ, Abdullah N (2005) Microstrip Sierpinski Carpet Antenna using Transmissiom Line Feeding. Microwave Conference Proceedings 2:900–903
8. Kliros GS, Liantzas KS, Konstantinidis AA (2007) Radiation Pattern Improvement of a Microstrip Patch Antenna using Electromagnetic Bandgap Substrate and Superstrate. WSEAS Transactions on Comunications 6:45–52
9. Sinha SN, Kumar PD (2003) Full wave analysis of scattering from a stub-loaded microstrip patch antenna. WSEAS Transactions on Communications 2:224–228
10. Huang J, Shan F, She J, Feng Z (2005) A Novel Small Fractal Patch Antenna. Microwave Conference Proceedings 4:780–783
11. Hamiti E, Ahma L, Sebak AR (2006) Computer Aided Design of U-Shaped Rectangular Patch Microstrip Antenna for Base Station Antennas of 900 MHz System. WSEAS Transactions on Communications 5:961–970
12. IE3D 10.0, Zeland Software Inc., Fremont, CA

Chapter 31

Reliability assessment and improvement of medium power induction motor winding insulation protection system using predictive analysis

M. Chafai, L.Refoufi, H. Bentarzi

DGEE, FSI, University of Boumerdes, Algeria, sisylab@yahoo.com

Abstract. This chapter presents a reliability assessment of a widely used protection system of medium-power squirrel cage induction motors. In conjunction with published field induction motors reliability data, this assessment effort is based on a predictive analysis integrating three predictive techniques: (1) a fault tree analysis (FTA) that allows to identify and then quantify the initiating events weighting factors; (2) an event tree analysis (ETA) that allows to predict the protection system probability outcomes following an external disturbance; and (3) a failure mode effect and criticality analysis (FMECA) that will help set the stage to develop a preventive maintenance program fit to keep up the induction motor protection system reliability at the required level with particular attention given to aggressive environmental factors such as found in cement plants.

Keywords. FTA, ETA, FMECA, Induction motors, Protection, Reliability, Failure modes

31.1 Introduction

The induction motor is the workhorse of industry. Despite its robustness and high reliability it has its physical limitations, which, if exceeded, will

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_31,

result in premature failure. Any operational failure will cause considerable economic losses. There is, therefore, a great need to improve the machine protection and hence its availability.

Dominant failure modes and failure mechanisms of some motor system parts, the initiating causes of motor failure, and their weighted contribution factors are first determined making use of published reliability data. Fault tree, event tree, and failure mode effects and criticality analyses are then developed for the assessment of the motor protection system reliability for further enhancement.

31.2 Induction motor stator winding failure mechanisms

Industrial surveys on machine reliability show [1] that the stator winding insulation is one of the most vulnerable components used in an AC electric machine. The failure of stator winding can be divided into

- insulation degradation and hence breakdown
- open circuit failure in the windings

31.2.1 Insulation failure mechanisms (IFM)

The stator winding insulation is always subjected to the combined thermal, electrical, mechanical, and environmental stresses during the long-term operation [2].

31.2.1.1 Thermal stress

Over time the insulation will deteriorate due to the normal thermal aging process; but the occurrence of premature failures, which are predominant, is a direct result of an overcurrent caused generally by an overload, a supply voltage imbalance, and/or voltage variations [2].

31.2.1.2 Electrical stresses

Most electrical failures are caused by a combination of overvoltage spikes and normal deterioration. This fast overvoltage can be caused by start-up switching, lightning, surges, and VFD to propagate through the material, the leading, to premature breakdown [4].

31.2.2 Winding wire open circuit failure

This failure, which rarely occurs, is generally due to the quality of wire as well as the level of electromechanical and environmental stresses pressed on the winding wire. The open circuit failure may occur at the terminal connections of the motor. The failure mechanism sequence of the induction motor is summarized in Fig. 31.1.

According to the statistical data given in Tables A.1 and A.2, the causes of failure of the motor insulation breakdown are predominant [1] and among them overload presents the highest percentage of causes followed by voltage imbalance and overvoltages as shown in Fig. 31.2.

The overload is mostly caused by mechanical problems due to excess loads or jams in the driven machine which forces the motor to develop higher torque, draw more current, and hence overheat [3].

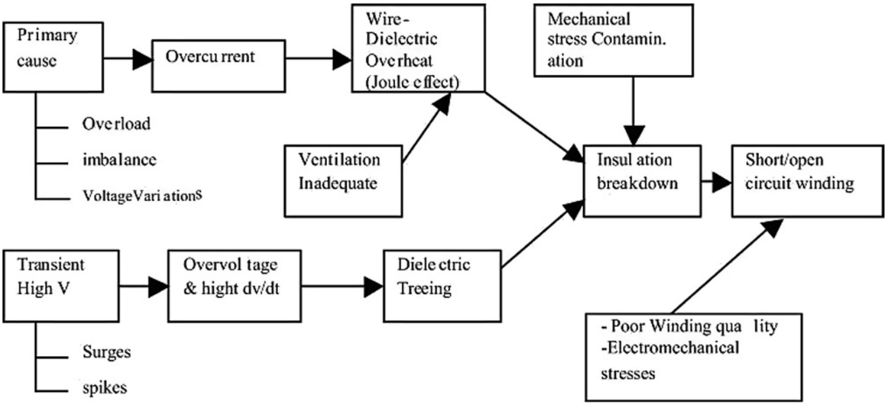


Fig. 31.1. Failure mechanism sequence of the electrical stator windings

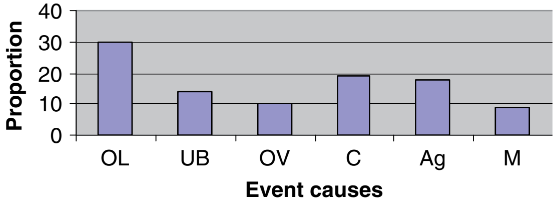


Fig. 31.2. Insulation failure initiating causes distribution

31.3 Failure probability quantification

Assuming that the failure rate of the motor is constant for a given time interval of 10^5 h and is evaluated as 10 F/10^6 h and that 40% of the motor failure is due to stator insulation breakdown then the probability of occurrence of the undesirable stator insulation breakdown is evaluated [5] as

$$F = 1 - R(t) = 1 - e^{-4.10^{-6}.t} \tag{31.1}$$

According to failure causes distribution, the contribution failure probability to the insulation breakdown of each initiating event (overload, voltage imbalance, etc.) is expressed as follows:

$$Fc = \alpha \,.\, F \tag{31.2}$$

The importance factor α and the contribution failure probabilities are quantified and given in Table 31.1.

Table 31.1. Insulation contribution failure probabilities

Initiating causes	Contribution factor α %	Fc	
OL (overload)	50	P_{OL}	2×10^{-2}
UB (single phasing)	20	P_{UB}	8×10^{-3}
OV (overvoltage)	10	P_{OV}	4×10^{-3}
OH (ambiant overheat)	10	P_{OH}	4×10^{-3}
Others	10	P_{OH}	4×10^{-3}

31.4 Fault tree analysis (FTA) of stator insulation failure

FTA is a top-down analysis technique that describes the relationship between basic causes, intermediate conditions, and top event such as motor breakdown [5]. The relationship is modeled in a tree-like structure with logical AND/OR gate as illustrated in Fig. 31.3. The numeric fault tree can then be used for determining the probability of the top event with the weighted importance factor of the causes. The quantitative evaluation will provide the priority protection parameter and determines the initiating events for further event tree analysis.

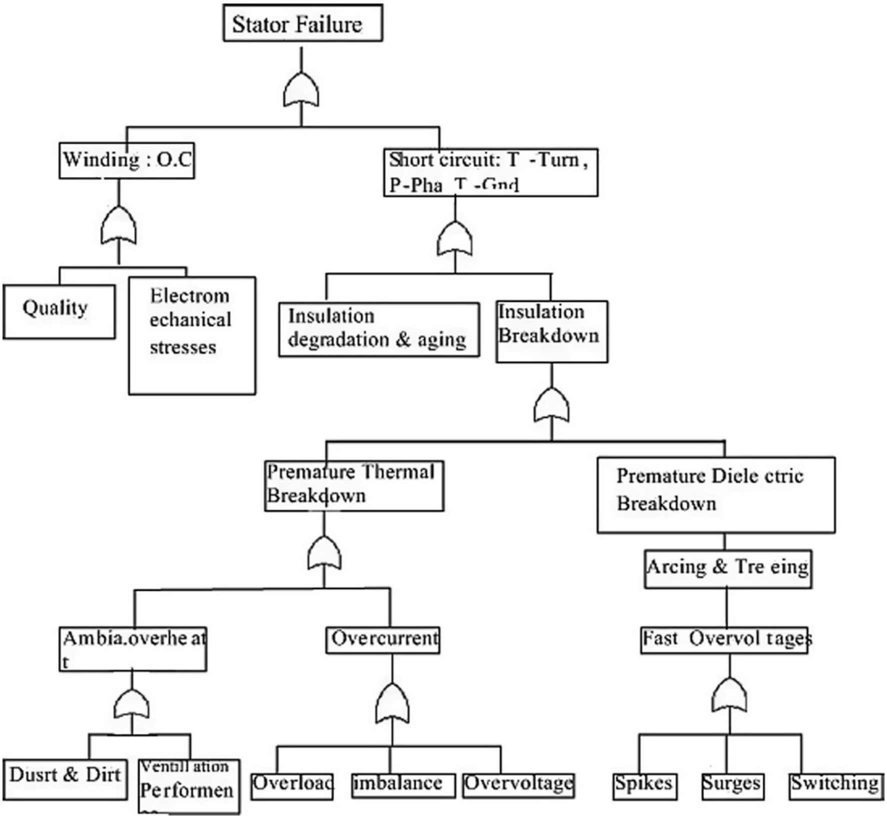


Fig. 31.3. Fault tree for the undesirable effect stator insulation failure

31.5 Protection system description

Based on the previous hierarchization of the initiating failure causes and IFM, the priority protection is provided first against overload followed by the unbalance (single phasing) and the overvoltages. The parameters to be controlled are mainly the current, the voltage, and the temperature [6].

While satisfying conditions such as discrimination, selectivity, and reliability, the credible optimized induction motor protection system consists basically of thermal relay, varistor, circuit breaker, fuses, and thermal sensor circuit as shown in Fig. 31.4.

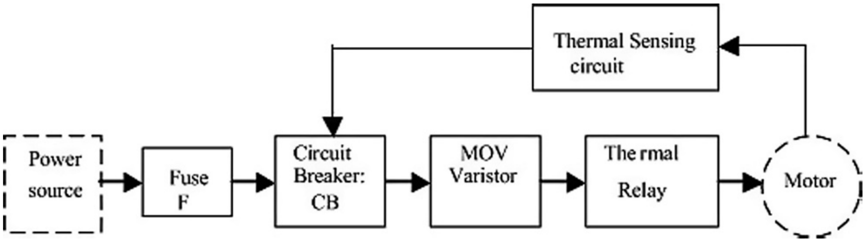


Fig. 31.4. Protection system

With the principle function of current and temperature detection–isolation, the thermal relay provides an overload protection. The metal oxide varistor (MOV) is used to clamp any slow or fast overvoltage from the power supply source. The circuit breaker is used as a switching device to protect the motor from short-circuit condition. The thermal sensor embedded on the stator winding is used to protect from high ambient temperature as well as the overload fault condition. The fuse opens its current responsive element in the case of an overcurrent or short-circuit condition.

The backup protection is provided in the case of the overload and single phasing. If the thermal relay fails to open, the thermal sensor circuit or the fuse is activated.

31.6 Event tree analysis (ETA)

An event tree starts with a specific initiating cause such as an overload, unbalance, or overvoltage as identified in the previous IFM and FTA and then follows the possible progression of the incident according to the success or failure of the protection devices.

This conducts to the elaboration of the sequence of events that leads to the insulation protection or severe motor failure and breakdown.

Each of the identified paths is evaluated [5]

- qualitatively by simplifying and eliminating impossible branches;
- quantitatively by attaching the probability to each event on the tree with the assumption that the failures are independent.
- The reduced and quantified ETAs for each initiating event are obtained from figs. 31.5, 31.6, 31.7 and 31.8.

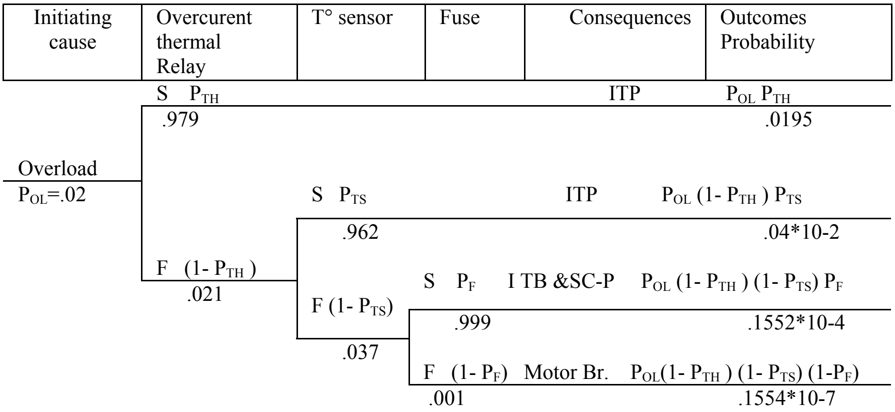


Fig. 31.5. Reduced ETA for the overload initiating event. (ITP: insulation thermal protection; ITB: insulation thermal breakdown)

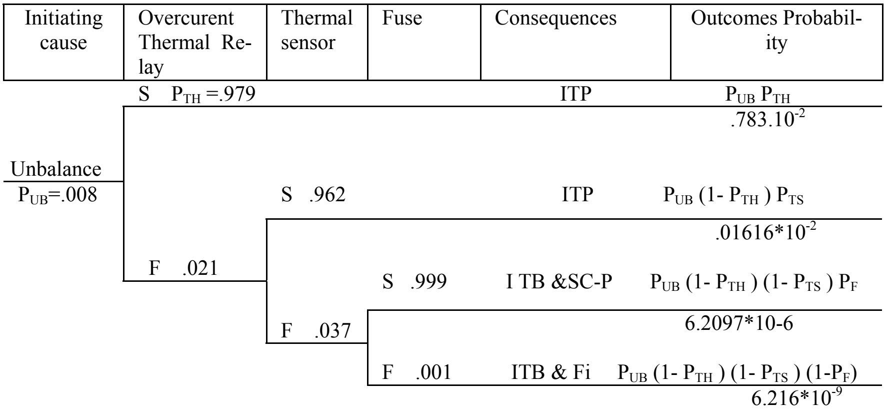


Fig. 31.6. ETA for a voltage unbalance (single phasing) initiating cause. (SC-P: short circuit protection)

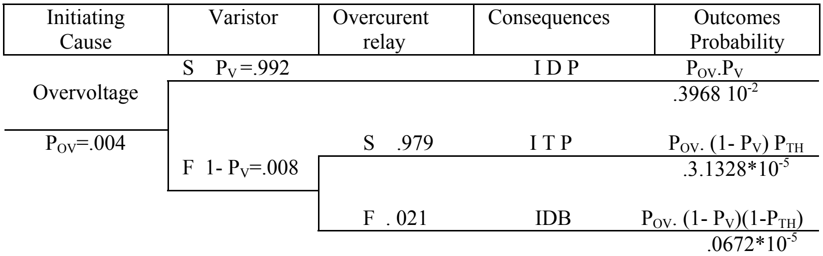


Fig. 31.7. ETA for initiating overvoltage event. (IDP: insulation dielectric breakdown)

Initiating cause	Thermal sensor Circuit	Consequences	Outcomes Probability
Ambient Oveheat P_{OH}=.004	S P_{TS} =.962	ITP	P_{OH}. P_{TS} =3848*10^{-2}
	F(1- P_{TS})=.037	I DB	P_{OH}. (1- P_{TS})=.0148*10^{-2}

Fig. 31.8. ETA for initiating ambient overheating event

The obtained results indicated in column 3 of Table A.4 show that the overall probability of insulation protection outcomes P_{IP} is much greater than that of the insulation breakdown outcome P_{IB} by more than 200 but further improvement is possible.

31.7 Protection system improvement

The improvement on the quality of the protection system will be based on an improved reliability and a continuous preventive maintenance of the protective devices to increase the probability of the insulation protection outcomes.

31.7.1 Reliability improvement

Improved reliability is obtained by the use of more reliable individual protective elements as well as by the redundancy (backup). This will increase the probability of success outcomes against dominant initiating events.

31.7.1.1 Better quality factor

The influence of quality factor [7] is shown according to part stress method relation [8] where the failure rate is given as

$$\lambda p = \lambda_B \ \pi_Q . \pi_E \tag{31.4}$$

where λ_B = base failure rate, π_Q = quality adjustment factor, and π_E = environment adjustment factor.

By using a better quality of the critical protective devices as shown in Table A.4, the obtained value of the consequences of insulation protection has increased while the probability of the insulation breakdown has decreased by a ratio of more than 2 as shown in Table A.4.

31.7.1.2 Redundancy

In this case, reliability can be increased by applying an active redundancy at thermal sensor circuit. A similar output circuit is added in parallel so that one can fail without causing system failure of the protection in the case of an ambient overheat or overload. The new increased reliability of the circuit is expressed as [9]

$$R_{\mathrm{Improved}} = 2R - R^2 \tag{31.5}$$

A substantial improvement is obtained

$$R_{\mathrm{Improved}}\,(= 0.998) > R_{Before}(= 0.962) \tag{31.6}$$

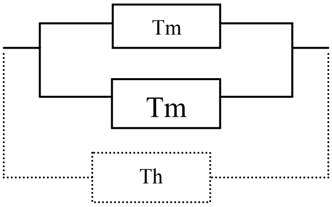


Fig. 31.9. Redundancy at thermal sensor circuit

Hence, thermal sensing circuit is a backup for thermal relay in the case of an overload condition and voltage imbalance as shown in Fig. 31.9.

31.7.2 Preventive maintenance on the protection system

To preserve an inherent reliability and successful function of this protection system a periodic preventive maintenance of its constitutive devices and their connection is required. Preventive actions and particularly environment protection against dust, temperature, vibrations, and contamination are taken in the light of FMECA.

The FMECA, which is an inductive method, seeks to identify the origin of potential failures and weak points in this protection system, classifies them in terms of criticality, and then determines the way of reducing their probability of occurrence in view of enhancing the reliability [5].

FMECA is developed in Table A.5 so that π_E factor is reduced and hence more than 20% reduction of the failure rate of the motor has been obtained [10].

This will prevent from any failure or degradation of the critical protective device and their connections leading to an undesirable event such as single phasing and loss of protection.

31.8 Conclusion

The assessment of an induction motor protection system reliability using fault tree and event tree analyses has been carried out. Field data-based calculations indicate that the probability of ensuring successful insulation protection is much greater than the occurrence of a failed insulation protection (leading to insulation breakdown) by a ratio of more than 200 as indicated in Table A.4.

Despite the high probability of successfully ensured insulation protection, a failure mode effects and criticality analysis clearly indicates that there is still a margin of improvement of system protection reliability. Through the selection of better quality protective devices, together with the use of redundancy where needed and a preventive maintenance program on the protection system proper in order to reduce the negative impacts of an aggressive environment such as that of cement plants, a drastic reduction of 60% in the probability of failed protection can be achieved.

Through the improvement of the motor protection system reliability, the motor reliability and availability can, therefore, be further improved and in a cost-effective manner.

Appendix

Table A.1. Motor failure statistics

Items	Failure (%)	
	[8]	[3]
Stator windings Ins.	37	30–40
Bearings	41	45–50
Rotor	10	8–2
Others	12	

Table A.2. Insulation failure causes

Failure causes	(%)
Overloads	30
Imbalance (single phasing and undervoltage)	14
Overvoltage	10
Contaminants	19
Aging	18
Miscellaneous	9

Table A.3. Failure rate and failure probability of protective devices [7] (B: before imporovement, A: after improvement)

Component	λb $F/10^6$ h	πE (GF)		λp		R		F	
		B	A	B	A	B	A	B	A
Over-I-Relay	0.25	1	1	2.04	0.68	0.979	0.993	0.021	0.007
Ckt Breaker	0.5	2	2	3	3	0.970	0.970	0.029	0.029
Fuse	0.010	2	2	0.02	0.02	0.999	0.999	0.001	0.001
Varistor	0.023	6	6	0.728	0.165	0.992	0.998	0.001	0.002
Thermal sensor	0.53	3	3	3.87	1.59	0.962	0.984	0.037	0.016

Table A.4. Probability of occurrence of the consequences. (IP: overall insulation protection, IB: insulation breakdown)

Initiating causes	Consequences	Probability		Ratio:
		Before	After improvement	A/B
OL	ITP	1.99155×10^{-2}	1.9998×10^{-2}	
UB	ITP	0.79976×10^{-2}	0.80826×10^{-2}	
OH	ITP	0.3848×10^{-2}	0.3936×10^{-2}	
OL+UB+OH	ITP $P_{itp}=\Sigma P_{itpi}$	3.17606×10^{-2}	3.20245×10^{-2}	
OV	IDP	$P_{dp} = 0.3968\times10^{-2}$	$0.3992.\times10^{-2}$	
OL+Ub+OV+OH	IP	$P_{IP} = 3.57286\times10^{-2}$	3.6016539×10^{-2}	
	IB	$P_{IB} = 1.48693\times10^{-4}$	0.640873×10^{-4}	0.43
			P_{IP} / P_{IB}	240

Table A.5. FMECA of protection system (C: criticality)

Item	Function	Failure mode	Cause	Effect	Preventive maintenance
Thermal relay	Overload protection	-Contacts fail shorted -Coil fails open -heater failure	-Contacts welded -Coil OC -Incorrect setting of tripping I	-Loss of thermal protection -Overheat	-Remove weld -Testability
Circuit breaker CB	Switching	-contacts fail shorted (stick occasionally) -contacts fail open -fails to active	-Contacts welded, corroded, and dirty -Mechanical failure binding -Incorrect setting of tripping I	-Loss of short circuit protection -Severe breakdown -Overload	-Cleaner vaporizer contacts - Dust removing - Trip setting
Fuses	Protection: SC- -overcurrent	-OC	-Overcurrent -Inadequate rating	-Shutdown -Unbalance	-Replacement -Adequate rating

(MOV)	Protection against fast overvoltage, surges, spikes	-SC -OC	-Excessive picks of voltage	-CB activated -Loss of dielectic protection	-Replacement -Adequate rating -Grounding filters
Thermal ckt	Stator ambient T° monitor	-OC	-Overheat dirty, dusty, and corrosive environment	-Overheat -Reduced aging	-Cleaning -Control ventilation
Terminals/wires	Electrical conduction	-Contacts fails -OC	-Loose screw -Poor contact -Corrosion	-Unbalance (single phase loss)	-Periodic check contact -Connections

References

1. O'Donnell P (1985) Report of large motor reliability survey of industrial and commercial installations. IEEE Transactions on Industry Applications IA-21(4):853–872
2. Bonett Austin H, Soukup George C (1992) Cause and analysis of stator and rotor failures in three-phase squirrel-cage induction motor. IEEE Transactions on Industry Applications 28(4):921–937
3. Bloch Heinz P, Geitner Fred K (1999) Machinery failure analysis and troubleshooting. Vol. 2, 3rd ed., New York: Elsevier
4. Curtis Lanham, President (2002) Understanding the tests that are recommended for electric motor predictive maintenance. New York: Baker Instrument Company, Energy publication
5. Ebeling Charles E (1997) An Introduction to reliability and maintainability engineering. New York: Ed Mc Graw–Hill
6. Wright A, Christopouls C (1993) Electrical power system protection. Boca Raton Chapman and Hall
7. Oraee Sharif H (1996) On-line protection machines stator windings against interturn insulation failures. University of Technology Iran *KEF* Industry Applicafions Magazine, pp 21–26
8. Military handbook MIL-HDBK-217 F (1991) Dept. of Defense (USA)
9. Chafai M, Refoufi L, Bentarzi H (2007) Medium induction motor winding insulation protection system reliability evaluation and improvement using predictive analysis. 6th WSEAS International Conference on Circuits and Sysems, Cairo, Egypt, pp 156–161
10. Rapport interne (2005) de suivi de maintenance et protection d'un moteur ventilateur des Cimenteries de Beni-saf et Chlef, Algeria

Chapter 32

Feature extraction by wavelet transforms to analyze the heart rate variability during two meditation techniques

G. Kheder, A. Kachouri, R. Taleb, M. ben Messaoud, M. Samet

Laboratory of Electronics and Technologies of Information, ENIS, B.P.W 3038 Sfax Tunisia, kheder_enis@yahoo.fr

Abstract. In this chapter, we present the analysis of HRV signals by wavelet transform. HRV, described by the extraction of the physiological rhythms embedded within its signal, is the tool through which adaptations of activity of the ANS have been widely studied. The assessment of wavelet transform (WT) as a feature extraction method was used in representing the electrophysiological signals. The purpose of all this is to study the ANS system of subjects who are doing meditation exercises such as the Chi and Yoga. The computed detail wavelet coefficients of the HRV signals were used as the feature vectors representing the signals. These parameters characterize the behavior of the ANS. In order to reduce the dimensionality of the data under study, the statistical parameters were computed.

Keywords. HRV, Wavelet, Feature extraction, ANOVA, Meditation

32.1 Introduction

In various countries, many people have died because of cardiac diseases. A great number of exams are conducted to obtain data of diversified nature, making subjects' visual evaluation harder. Added to that, visual fatigue is the main cause making the manual analysis error prone. Hence, it

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_32,

is suitable to develop an automatic system to process the electrocardiogram (ECG) signal. In order to allow an early diagnosis and an efficient treatment, the recognition of patterns of cardiac diseases can be improved through automatic feature extraction [1–3]. An electrocardiogram (ECG) can be defined as an electrical signal that represents the heart's cardiac activity. In general, this signal is recorded by means of a certain number of electrodes which are pasted on the body. The most important waves responsible for the formation of a typical ECG are generally the P, QRS, and T waves. The P wave corresponds to the atrium's depolarization. The QRS complex results from the ventricular depolarization. The T wave corresponds to the polarization of the ventricle.

In general, the signal, which corresponds to the atrium polarization, is merged with the QRS one. What is noticed here is that the ECG beat's shape can dynamically change. It is also highly correlated with the pathology type. The R-wave that manifests the depolarization process of the ventricle is the largest amplitude of a single cycle of the normal ECG [2, 4].

RR interval is the time between successive R-waves and RR tachogram is the series of RR intervals. Thus, in this time series, variability has been largely used as a measure of the heart's function. Through this we can identify risky patients who are prone to cardiovascular events or death. In fact, in this time series, variations' analysis is known as heart rate variability (HRV) analysis [5, 6].

The parameter used in assessing autonomous nervous system (ANS) activity is defined as heart rate variability. HRV, described by the extraction of the physiological rhythms embedded within its signal, is the tool through which adaptations of activity of the ANS have been widely studied. The HRV's non-stability presents a challenge to the technical aspects of its measurement, especially in the dynamic conditions of functional testing [7].

There have been various mathematical methods to analyze HRV. The most common one is the Fourier transform, but it is limited to stationary signals. The most important thing to do, while calculating a signal expansion, is to localize a given basis function in time and in frequency. For instance, in Fourier transform, while analyzing, the functions used are infinitely sharp in their frequency localization. They exist at one exact frequency but have no time localization due to their infinite extends [8].

In order to define a particular basis function's localization, we can find different ways but they are linked to the expansion of the function in time and frequency. In fact, to overcome this very limitation, we applied the wavelet transform (WT). Wavelet analysis is one among the options available that may help to quantify HRV in non-stationary conditions [7].

Wavelet transform (WT) represents a mathematical way to study nonstationary signals. Therefore, its usefulness has been increasingly adapted over the last 10 years. It was employed in different fields such as communication technology, geophysics, and image processing.

32.2 Methods

The RR interval variations present during resting conditions represent a fine-tuning of beat-to-beat control mechanisms. Because it helps to evaluate the equilibrium between the sympathetic and parasympathetic influences on heart rhythm, HRV signal analysis is very important and crucial for the study of the autonomic nervous system (ANS). The nervous system's sympathetic branch increases the heart rhythm resulting in shorter beat intervals whereas the parasympathetic branch decelerates the heart rhythm leading to longer beat intervals. The spectral analysis of the HRV has led to the identification of two fairly distinct peaks: high (0.15–0.5 Hz) and low (0.05–0.15 Hz) frequency bands. Fluctuations in the heart rate, occurring at the spectral frequency band of 0.15–0.5 Hz, known as high-frequency (HF) band, reflect parasympathetic (vagal) activity, while fluctuations in the spectral band (0.05–0.15 Hz), known as low-frequency (LF) band, are linked to the sympathetic modulation, but includes some parasympathetic influence (sympathetic–vagal influences) [6]. It is now established that the level of physical activity is clearly indicated in the HRV power spectrum. For example, when a healthy subject stands up there is an increase of HRV in the LF spectral band, which is considered an estimate of the sympathetic influence on the heart. Consequently, the LF/HF ratio is considered to mirror sympathovagal balance or to reflect sympathetic modulations [8–10].

In this very research, we are looking for an effective way to analyze the HRV with advanced technique of signal processing. The purpose of all this is to study the ANS system of subjects who are doing meditation exercises such as the Chi and Yoga.

32.3 Justification and purpose of the study

The purpose of this study is the separation of the two bands of frequency: HF and LF of HRV signal through the multiresolution decomposition by discrete wavelet transform. After the access to these components, which inform us about the function of ANS system, we can demonstrate that the

WT analysis can be an effective clinical tool in examining the heart rhythm.

This proposition will be applied on an article drawn from physioBank database [www.physionet.org] entitled "Exaggerated heart rate oscillations during two meditation techniques," which contains a data of heart rate time series of two groups of healthy subjects experiencing two series of records, one as premeditation and the other during the period of meditation (during specific traditional forms of Chinese Chi and Kundalini Yoga meditations). We have used another control signal for sane subjects in normal respiration (Table 32.1).

These data were used by [9] who applied both spectral analysis and analytic technique based on the Hilbert transform to quantify the heart rate. This method proved to be not very effective compared to the wavelet method.

The ability of this latter technique to give simultaneous time and frequency resolutions and separation of sympathetic and parasympathetic bands makes it an ideal tool for studying HRV.

Table 32.1. Database

Exercise	Notation
Chi meditation	C1, C2,…, C8 (before and during)
Yoga meditation	Y1, ….,Y4 (before and during)
Normal respiration	N1, N2,…, N11

32.4 Feature extraction

32.4.1 The proposed feature extraction methodology

By means of wavelet analysis, a matrix of data is obtained, where time and frequency domain information is present. A mother waveform is "compressed" or "stretched" to obtain wavelets of different scales that are used along time comparing them with the original signal. Low-scale levels correspond to rapidly changing details or high frequency, whereas high-scale levels correspond to slow changing details or low frequency. For every scale level and time a correlation coefficient was obtained, representing the correspondence between the analysis wavelet and the original signal. For example, a high correlation coefficient between the original signal and a low-scale wavelet at the beginning of the record means that high-frequency components are present at that time. Thus, this coefficient provides information about the moment that the RR interval is changing (time domain) and about the frequencies that are involved in these changes (frequency domain). In short-term recordings, high-frequency (HF) components (0.15–0.40 Hz) reflect vagal activity, while low-frequency (LF) components (0.04–0.15 Hz) are considered to be under the influence of both sympathetic and parasympathetic tone. Thus an increased LF/HF ratio may indicate either increased sympathetic activity or decreased vagal tone [32, 13].

Tests are often carried out using different types of wavelets and the most effective one is chosen for the particular application. We can detect changes of the HRV signals by means of the smoothing feature of Daubechies wavelet of order (db4). This wavelet (db4) is used by most researchers [7–10, 12, 13] to analyze the HRV. Therefore, the wavelet coefficients were computed using db4 in the present study. The wavelet coefficients were computed using the MATLAB software package.

Selection of appropriate wavelets and the number of decomposition levels are very important in the analysis of signals using the WT. The number of decomposition levels is chosen based on the dominant frequency components of the signal. The levels are chosen such that those parts of the signal that correlate well with the frequencies required for classification of the signal are retained in the wavelet coefficients. In order to determine the appropriate number of decomposition levels, different experiments were performed. In the present study, the number of decomposition levels was chosen to be 6. The HRV signal was resampled at 4 Hz. Thus, scale 1 corresponds to 2–1 Hz, and scale 2 to 1–0.5 Hz. Scales 3 and 4 correspond approximately to HF (0.125–0.5 Hz), and scales 5 and 6 to LF (0.03125–0.125 Hz).

Table 32.2. Localization of the two bands LF and HF after decomposition by DWT

HRV component	Scales	Frequency bands (Hz)
	D1	1–2
	D2	0.5–1
HF	D3	0.25–0.5
	D4	0.125–0.25
LF	D5	0.0625–0.125
	D6	0.03125–0.0625

The computed detail wavelet coefficients of the HRV signals were used as the feature vectors representing the signals. These parameters characterize the behavior of the ANS. In order to reduce the dimensionality of the data under study, the following statistical parameters were computed:

- STDLF and STDHF: Standard deviation of the wavelet coefficients in each sub-band.
- %LF and %HF: LF and HF powers of wavelet coefficients in each sub-band measured in normalized units:
- %LF = LF/(LF+HF)×100; %HF = HF/(LF+HF)×100.
- R: ratio; R=LF/HF

32.4.2 Statistics analysis

The ANOVA test takes into account not only changes in mean values and standard deviation but also changes occurring in each subject in different exercises and meditations. This test is considered significant when $p < 0.05$, where p is the probability of the data recorded from the same subject in the two states before and during the meditation.

32.5 Results and discussion

Figure 32.1 shows the parameters taken from the wavelet coefficients of the HRV signals of three subjects: two during meditation exercises and the third subject in normal respiration state who is referred to as control in the histogram.

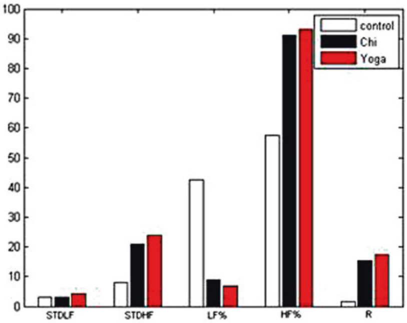


Fig. 32.1. Parameters extracted from HRV before and during Chi and Yoga meditation and control group

We notice that the ANS behavior is characterized by a decrease of the concentrated power in the LF band and an increase at the HF band level.

During the tests, the pursuit of these changes are clear in the LF and HF powers of wavelet coefficient in each sub-band measured in normalized units (LF% and HF%).

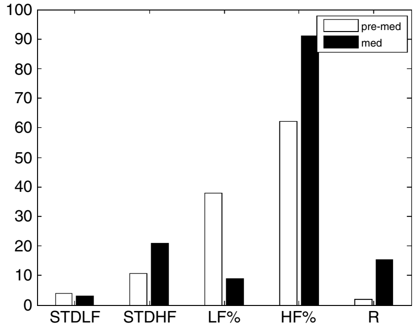


Fig. 32.2. Parameters extracted from HRV before and during Chi meditation

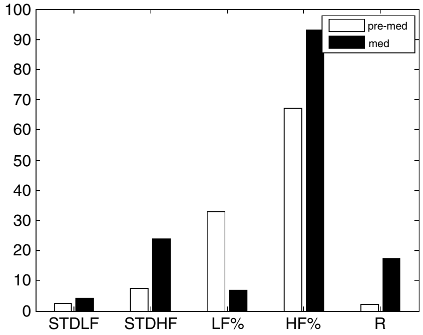


Fig. 32.3. Parameters extracted from HRV before and during Yoga meditation

The results depicted in Figs. 32.2 and 32.3 demonstrate that the percentage of HF during the meditation is greater than the percentage of HF in the premeditation, which indicates that the sympathetic nerves are more active during meditation and this situation causes the heart rate of the subjects to be quicker than the ordinary situation.

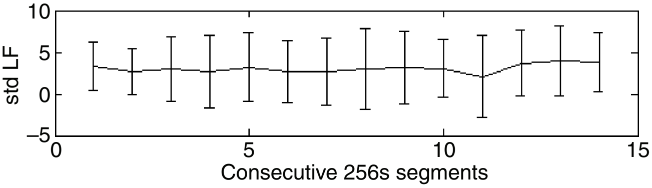


Fig. 32.4. Variation of standard deviation of wavelet coefficients in LF band over each consecutive drive segment

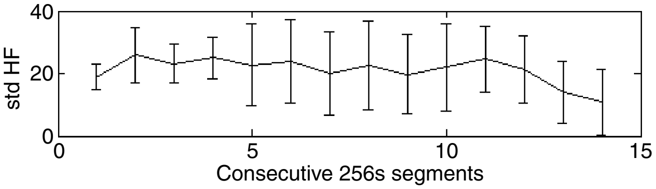


Fig. 32.5. Variation of standard deviation of wavelet coefficients in HF band over each consecutive drive segment

Studying the different segments of 256s, we can notice that the peak of the high variation indexed on Figs. 32.4 and 32.5 error bars presents a relative stability on standard deviation of LF. Moreover, the peak that represents the acceleration of respiration rhythm is localized on the variation of standard deviation HF.

Table 32.3. p-Value: statistics test

Parameters	Chi	Yoga
STDLF	$p < 0.0007$	$p < 0.0072$
STDHF	$p < 8.1154e\text{-}008$	$p < 0.0003$
%LF and %HF	$p < 5.6301e\text{-}007$	$p < 7.5352e\text{-}006$
R	$p < 1.1877e\text{-}006$	$p < 0.0287$

The statistical test ANOVA on the subject before and during the meditation is always significant ($p < 0.05$). However, when we increase the data base, we move to nonsignificant values of the balance of LF/HF.

32.6 Conclusion

Using wavelet transforms, we managed to separate the two essential components (LF and HF) of the HRV signal. The analysis of the two bands LF and HF described by wavelet coefficients informs us about the behavior of ANS. The response of ANS during Chi and Yoga was studied well in time–frequency domain by calculation of statistical parameters and the localization of energy concentration reflect as a result of amplitude variation, that is, the acceleration and the refraining of sympathetic and parasympathetic systems.

Thus we can deal with the approaches of classification with a guarantee of ample feasibility of anomalies of cardiac insufficiency.

References

1. Madeiro PV, Cortez Paulo C, Oliveira Francisco I, Siqueira Robson S (2007) A new approach to QRS segmentation based on wavelet bases and adaptive threshold technique. Medical Engineering & Physics 29:26–37
2. El Khansa L, Naït-Ali A (2007) Parametrical modelling of a premature ventricular contraction ECG beat: Comparison with the normal case. Computers in Biology and Medicine 37:1–7
3. Engin M (2006) Spectral and wavelet based assessment of congestive heart-failure patients. Computers in Biology and Medicine
4. Khoo MCK, Kim T, Berry RB (1999) Spectral Indices of Cardiac Autonomic Function in Obstructive Sleep Apnea. SLEEP 22(4)
5. Faust O, Acharya R, Krishnan SM, Min LC (2004) Analysis of cardiac signal using spatial filling index and time-frequency domain. BioMedical Engineering OnLine
6. Jafarnia-Dabanlooa N, McLernona DC, Zhangb H, Ayatollahic A, Johari-Majd V (2007) A modified Zeeman model for producing HRV signals and its application to ECG signal generation. Journal of Theoretical Biology 244:180–189
7. Pichot V, Gaspoz JM, Molliex S, Antoniadis A, Busso T, Roche F, Costes F, Quintin L, Lacor JR, Barthelemy J (1999) Wavelet transform to quantify heart rate variability and to assess its instantaneous changes. Journal of Applied Physiology 86:1081–1091
8. Belova NY, Mihaylov SV, Piryova G (2007) Wavelet transform: A better approach for the evaluation of instantaneous changes in heart rate variability. Autonomic Neuroscience: Basic and Clinical 131:107–122
9. Jou-Wei Lin, Juey-Jen Hwang, Liang-Yu Lin, Jiunn-Lee Lin (2006) Measuring Heart Rate Variability with Wavelet Thresholds and Energy Components in Healthy Subjects and Patients with Congestive Heart Failure. Cardiology 106:207–214
10. Vigo DE, Guinjoan SM, Scaramal M, Siri LN, Cardinali DP (2005) Wavelet transform shows age-related changes of heart rate variability within independent frequency components. Autonomic Neuroscience: Basic and Clinical 123:94–100
11. Peng CK, Mietus JE, Liu Y, Khalsa G, Douglas PS, Benson H, Goldberger AL (1999) Exaggerated Heart Rate Oscillations During Two Meditation Techniques. International Journal of Cardiology 70:101–107
12. Toledo E, Gurevitz O, Hod H, Eldar M, Akselro S (2003) Wavelet analysis of instantaneous heart rate: a study of autonomic control during thrombolysis. American Journal of Physiology Regulatory Integrative and Comparative Physilology 284:1079–1091
13. Burri H, Chevalier P, Arzi M, Rubel P, Kirkorian G, Touboul P (2006) Wavelet transform for analysis of heart rate variability preceding ventricular arrhythmias in patients with ischemic heart disease. International Journal of Cardiology 109:101–107

Chapter 33

Fractional mechanical model for the dynamics of non-local continuum

G. Cottone, M. Di Paola, M. Zingales

Dipartimento di Ingegneria Strutturale e Geotecnica University of Palermo, Viale delle Scienze 90128, Palermo, Italy
giuliocottone@diseg.unipa.it; dipaola@diseg.unipa.it; zingales@diseg.unipa.it

Abstract. In this chapter, fractional calculus has been used to account for long-range interactions between material particles. Cohesive forces have been assumed decaying with inverse power law of the absolute distance that yields, as limiting case, an ordinary, fractional differential equation. It is shown that the proposed mathematical formulation is related to a discrete, point-spring model that includes non-local interactions by non-adjacent particles with linear springs with distance-decaying stiffness. Boundary conditions associated to the model coalesce with the well-known kinematic and static constraints and they do not run into divergent behavior. Dynamic analysis has been conducted and both model shapes and natural frequency of the non-local systems are then studied.

Keywords. Non-local elasticity, Fractional calculus, Power law attenuation function, Modes of vibration and dynamics of non-local bar

33.1 Introduction

Non-local continuum mechanics has received a growing interest in the late 1960s as reported in several studies [21,22,15], since it is able to describe

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_33,

the microstructural behavior of materials. The use of these theories explains some phenomena that may be unpredicted by classical theory of local continuum mechanics; for instance, the unrealistic stress singularities at crack tips are smoothed by a non-local approach.

Non-local continuum mechanics has been treated with two different approaches: The *gradient elasticity theory* (weak non-locality) and the *integral non-local theory* (strong non-locality). The first approach mainly consists in the introduction of gradient strains in the constitutive equations [27,1]. The main drawback of gradient elasticity model regards the fulfillment of boundary conditions, even though it is to be mentioned that some strategies for overcoming this drawback has recently been proposed [32,33,34,7]. The non-local integral model has been introduced as intuitive extension of interpolation formulas of molecular dynamics [21,15]. The non-local interaction is represented as a convolution integral in which the kernel is a decaying function with the inter-distance of different points. In this setting, several papers are available [4,5,30,6,16,17].

Recently, the problem of non-local continuum has been faced by fractional calculus approach [23,14]. The approach in terms of fractional calculus is an intermediate one between gradient and integral formulation and then it is very attractive from a conceptual point of view. In particular, it has been show in [14] that the long-range interaction may be cast in terms of Marchaud fractional derivatives for an infinite domain, while it remains only the integral part of the Marchaud fractional derivative for a bounded domain.

In this chapter, the problem is first formulated in terms of the total energy stored showing that directly postulating the form of the total energy stored, some inconsistencies appear. Then, a consistent physical model is presented and the main result shown is that the governing equations are fractional differential equations. Dynamics of the non-local continuum follows, therefore, straightforwardly from the physical model taking into account the inertial forces.

33.2 Basics on fractional calculus

The well-established theory of fractional differentiation deals with derivatives and integrals of any real (or even complex) order. Although this theory is as old as the common differential calculus, it has not gained interest from engineers and physicists until the end of the last century, with the appearance of monographs and conferences on this topic. Further, the lack of an easy geometrical meaning of the derivative of real order causes a diffuse wariness also nowadays. In fact, while we are aware of the geometrical meaning of the first or second derivative of a function, for example, the meaning of the derivative of order $1/2$ is a task that has not been solved for 300 years. Actually, it was already L'Hôpital who asked Leibniz "What if n (in $\mathrm{d}^n f/\mathrm{d}x^n$) be $1/2$?" in 1695!

Other difficulties that arise in dealing with fractional calculus is that there are many definitions that generalize the ordinary calculus and, further, the calculations involved are always hard to be tackled by hand. Despite these difficulties, the application of the fractional calculus asserts to reach very interesting results. For instance, interesting applications can be found in physics and biophysics [19], in polymer rheology [37], in fracture mechanics [8–10], and in viscoelasticity [3]. In stochastic dynamics, we report the interesting analysis of linear and non-linear systems driven by fractional Brownian motion [24–25] or driven by Lévy α-stable white noise processes [11, 18], whose probability density of the response is ruled by a fractional differential equation, involving fractional derivative in the diffusive term. In probability theory, a representation of the statistics of random variables by means of the fractional calculus has been recently proposed [12] and the path integral solution has been reformulated in terms of fractional moments in order to solve stochastic differential equations [13].

In the following, we report the main definitions of the most important fractional integrals and derivatives highlighting the properties used in the definition of non-local elasticity. For an exhaustive treatment of the topic and for rigorous proofs, the reader is referred to the excellent encyclopedic monographs of Samko, et al. [35] or to [20, 26, 29, 31].

Many are the names of those prominent mathematicians of the past who contributed to the theory of fractional calculus: Riemann, Euler, Laplace, Fourier, Abel, Liouville, and Weyl among others played an important role in what now is the corpus of the fractional calculus. Many are also the fractional operators in literature causing a great effort needed to approach the theory. In the spirit of this chapter, we present only the Riemann–Liouville (RL) fractional integral and derivative and the Marchaud (Ma) fractional derivative.

33.2.1 The Riemann–Liouville fractional integral and derivative

First of all, it must be highlighted that the so-called "*fractional*" calculus is a misnomer, because actually it deals with the *generalization of the differential calculus* to real or complex order. By the way, this name 408 is consolidated among mathematicians and scientists in many fields and is kept for historical reasons.

A simple approach to introduce the fractional operators is constituted by three steps: (i) consider first the *n*-folding operation on a definite integral; (ii) extend the folding operation to a real number of folding, obtaining then, what is called the fractional integral; (iii) prove that the fractional integral has an inverse operator, called the fractional derivative. Other approaches to the definition of the fractional operators are reported in the book of Miller and Ross [26].

In this first part, we consider functions defined in a finite interval. Given a Lebesgue measurable function $f(x)$ on the closed interval $[a,b]$, briefly $f(x) \in Leb_1([a,b])$, according to the notation in [35], we indicate with $\left(I_{a+}f\right)(x)$ the following definite integral:

$$\left(I_{a+}f\right)(x) \overset{def}{=} \int_a^x f(\xi)\,\mathrm{d}\xi, \qquad x > a \tag{33.1}$$

Integrating twice, the resulting function will be indicated as $\left(I_{a+}^2 f\right)(x)$, and it can be recast in the form

$$\left(I_{a+}^2 f\right)(x) = \int_a^x \int_a^{\xi_1} f(\xi_2)\,\mathrm{d}\xi_1\,\mathrm{d}\xi_2 = \int_a^x \int_{\xi_1}^x f(\xi_1)\,\mathrm{d}\xi_2\,\mathrm{d}\xi_1 =$$

$$= \int_a^x f(\xi_1) \int_{\xi_1}^x \mathrm{d}\xi_2\,\mathrm{d}\xi_1 = \int_a^x f(\xi_1)(x-\xi_1)\,\mathrm{d}\xi_1$$

A further integration leads to

$$\left(I_{a+}^3 f\right)(x) = \underbrace{\int_a^x \mathrm{d}x \cdots \int_a^x f(\xi)\,\mathrm{d}\xi}_{3-fold} = \frac{1}{2}\int_a^x (x-\xi)^2 f(\xi)\,\mathrm{d}\xi$$

and for a generic integer number of folding, $n \in \mathbb{N}$, the well-known Cauchy equation

$$\left(I_{a+}^n f\right)(x) = \underbrace{\int_a^x \mathrm{d}x \cdots \int_a^x f(\xi)\,\mathrm{d}\xi}_{n-fold} = \frac{1}{(n-1)!}\int_a^x (x-\xi)^{n-1} f(\xi)\,\mathrm{d}\xi \tag{33.2}$$

holds. Calling γ a positive real number, the generalization of $\left(I_{a+}^n f\right)(x)$, written for an integer number of folding operations, into $\left(I_{a+}^\gamma f\right)(x)$ can be found from the latter equation by means of the Euler gamma function that is defined as the integral

$$\Gamma(\gamma) \overset{def}{=} \int_0^\infty \xi^{\gamma-1} \exp(-\xi)\,\mathrm{d}\xi \tag{33.3}$$

and interpolates the factorial function, that is, $(n-1)! = \Gamma(n)$. Then, one obtains from Eq. (33.2)

$$\left(I_{a+}^{\gamma} f\right)(x) \overset{def}{=} \frac{1}{\Gamma(\gamma)} \int_a^x \frac{f(\xi)d\xi}{(x-\xi)^{1-\gamma}}, \text{ with } \gamma > 0 \tag{33.4}$$

called *left-sided Riemann–Liouville fractional integral* of real order γ. Analogously, given $f(x) \in Leb_1([a,b])$, the same procedure can be applied to the integral with fixed upper limit

$$\left(I_{b-} f\right)(x) \overset{def}{=} \int_x^b f(\xi)d\xi, \qquad x < b \tag{33.5}$$

leading to the operator

$$\left(I_{b-}^{\gamma} f\right)(x) \overset{def}{=} \frac{1}{\Gamma(\gamma)} \int_x^b \frac{f(\xi)d\xi}{(\xi-x)^{1-\gamma}}, \qquad \gamma > 0 \tag{33.6}$$

called *right-sided Riemann–Liouville fractional integral.* Now, in order to define an inverse operator of Eqs. (33.4) and (33.6), we shall prove that this relation is invertible in the sense that for a given function $\varphi(x)$, there exists a function $f(x)$, solution of the integral equation

$$\varphi(x) = \frac{1}{\Gamma(\gamma)} \int_a^x \frac{f(\xi)d\xi}{(x-\xi)^{1-\gamma}}, \qquad x > a \tag{33.7}$$

This equation has been solved by Abel, for $0 < \gamma < 1$, who found that its solution is unique and given by

$$f(x) = \frac{1}{\Gamma(1-\gamma)} \frac{d}{dx} \int_a^x \frac{\varphi(\xi)d\xi}{(x-\xi)^{\gamma}} \tag{33.8}$$

In the same way, the solution of the integral equation consequent to the definition in Eq. (33.6) is given in the form

$$\varphi(x) = \frac{1}{\Gamma(\gamma)} \int_x^b \frac{f(\xi)d\xi}{(\xi-x)^{1-\gamma}}, \qquad x \le b \tag{33.9}$$

and its solution has been proved to be

$$f(x)=-\frac{1}{\Gamma(1-\gamma)}\frac{d}{dx}\int_x^b\frac{\varphi(\xi)d\xi}{(\xi-x)^\gamma},\qquad 0<\gamma<1 \tag{33.10}$$

Therefore, it is possible to define the *left-handed RL fractional derivative*, $\left(\mathcal{D}_{a+}^\gamma f\right)(x)$, given by

$$\left(\mathcal{D}_{a+}^\gamma f\right)(x)\overset{def}{=}\frac{1}{\Gamma(1-\gamma)}\frac{d}{dx}\int_a^x\frac{f(\xi)d\xi}{(x-\xi)^\gamma},\qquad 0<\gamma<1 \tag{33.11}$$

and the *right-handed RL fractional derivative*, $\left(\mathcal{D}_{b-}^\gamma f\right)(x)$, in the form

$$\left(\mathcal{D}_{b-}^\gamma f\right)(x)=\frac{(-1)}{\Gamma(1-\gamma)}\frac{d}{dx}\int_x^b\frac{f(\xi)dt}{(\xi-x)^\gamma},\qquad 0<\gamma<1 \tag{33.12}$$

Useful representations of Eqs. (33.11) and (33.12) are

$$\left(\mathcal{D}_{a+}^\gamma f\right)(x)=\frac{1}{\Gamma(1-\gamma)}\left[\frac{f(a)}{(x-a)^\gamma}+\int_a^x\frac{f'(\xi)d\xi}{(x-\xi)^\gamma}\right],\qquad 0<\gamma<1 \tag{33.13}$$

and

$$\left(\mathcal{D}_{b-}^\gamma f\right)(x)=\frac{1}{\Gamma(1-\gamma)}\left[\frac{f(b)}{(b-x)^\gamma}-\int_x^b\frac{f'(\xi)d\xi}{(\xi-x)^\gamma}\right],\qquad 0<\gamma<1 \tag{33.14}$$

as reported in [35, Eqs. (2.24 and 2.25)].

In order to extend the definition of fractional derivative of order greater than 1, first we recall a standard notation, indicating with $[\gamma]$ the integer part of a real, number and with $\{\gamma\}$ the fractional part, that is, $\gamma=[\gamma]+\{\gamma\}$. Then, for every positive real number γ, the Riemann–Liouville fractional derivatives are defined as

$$\left(\mathcal{D}_{a+}^\gamma f\right)(x)=\frac{1}{\Gamma(n-\gamma)}\frac{d^n}{dx^n}\int_a^x\frac{f(\xi)dt}{(x-\xi)^{\gamma-n+1}},\qquad n=[\gamma]+1 \tag{33.15}$$

$$\left(\mathcal{D}_{b-}^\gamma f\right)(x)=\frac{(-1)^n}{\Gamma(n-\gamma)}\frac{d^n}{dx^n}\int_x^b\frac{f(\xi)d\xi}{(\xi-x)^{\gamma-n+1}},\qquad n=[\gamma]+1 \tag{33.16}$$

Comparing the definitions, it follows that the fractional derivatives and fractional integrals are related by the simple relations

$$\left(\mathcal{D}_{a+}^{\gamma} f\right)(x) = \frac{d^n}{dx^n}\left(I_{a+}^{n-\gamma} f\right)(x), \qquad n = [\gamma] + 1 \tag{33.17}$$

$$\left(\mathcal{D}_{b-}^{\gamma} f\right)(x) = (-1)^n \frac{d^n}{dx^n}\left(I_{b-}^{n-\gamma} f\right)(x), \qquad n = [\gamma] + 1 \tag{33.18}$$

Generalization of Eq. (33.13) for every $\gamma > 0$ [35, Eq. (2.43)] is

$$\left(\mathcal{D}_{a+}^{\gamma} f\right)(x) = \sum_{k=0}^{n-1} \frac{1}{\Gamma(1+k-\gamma)} \frac{f^{(k)}(a)}{(x-a)^{\gamma-k}} + \frac{1}{\Gamma(n-\gamma)} \int_a^x \frac{f^{(n)}(\xi)\,d\xi}{(x-\xi)^{\gamma-n+1}} \tag{33.19}$$

and of Eq. (33.14) is

$$\left(\mathcal{D}_{b-}^{\gamma} f\right)(x) = \sum_{k=0}^{n-1} \frac{1}{\Gamma(1+k-\gamma)} \frac{f^{(k)}(b)}{(b-x)^{\gamma-k}} + \frac{(-1)^n}{\Gamma(n-\gamma)} \int_a^x \frac{f^{(n)}(\xi)\,d\xi}{(\xi-x)^{\gamma-n+1}} \tag{33.20}$$

derived by direct calculations.

The presence of the derivatives of order *n* in the fractional derivatives' definitions involves more strict conditions ([35], p. 37) to the existence of the fractional derivative. A sufficient condition is the function having continuous derivatives up to the order $[\alpha]-1$.

The definitions of fractional operators on the interval $[a,b]$ are easily extended to the case of the half-axis or the whole axis, obtaining another class of derivatives, in some literature [26] indicated as *Liouville–Weyl*. We prefer to follow the indication of [35], calling them as Riemann–Liouville. Letting the extremes of the interval going to infinity, in particular it is obtained as

$$\left(I_{+}^{\gamma} f\right)(x) \overset{def}{=} \frac{1}{\Gamma(\gamma)} \int_{-\infty}^{x} \frac{f(\xi)\,d\xi}{(x-\xi)^{1-\gamma}}, \qquad \gamma > 0,\ x \in \mathbb{R} \tag{33.21}$$

$$\left(I_{-}^{\gamma} f\right)(x) \overset{def}{=} \frac{1}{\Gamma(\gamma)} \int_{x}^{\infty} \frac{f(\xi)}{(\xi-x)^{1-\gamma}}\,d\xi, \qquad \gamma > 0,\ x \in \mathbb{R} \tag{33.22}$$

that can be expressed, with a change of variables, in a more compact way

$$\left(I_{\pm}^{\gamma} f\right)(x) \overset{def}{=} \frac{1}{\Gamma(\gamma)} \int_{0}^{\infty} \frac{f(x \mp \xi)}{\xi^{1-\gamma}}\,d\xi, \qquad \gamma > 0,\ x \in \mathbb{R} \tag{33.23}$$

The *Riemann–Liouville fractional derivative* descends from Eq. (33.23) and assumes the form

$$\left(\mathcal{D}_{\pm}^{\gamma} f\right)(x)=\frac{1}{\Gamma(1-\gamma)} \frac{d}{dx} \int_0^{\infty} \frac{f(x \mp \xi)\, d\xi}{\xi^{\gamma}} \tag{33.24}$$

for $0<\gamma<1$ or, for $\gamma>0$, it is expressed as

$$\left(\mathcal{D}_{\pm}^{\gamma} f\right)(x)=\frac{(\pm 1)^n}{\Gamma(n-\gamma)} \frac{d^n}{dx^n} \int_0^{\infty} \xi^{n-\gamma-1} f(x \mp \xi)\, d\xi, \quad n=[\gamma]+1 \tag{33.25}$$

33.2.2 The Marchaud definition

On the real axis Eq. (33.24) can be written in a more convenient form, working out a little on definition as reported in [35]. In fact, suppose first that the function $f(x)$ is continuously differentiable and that its first derivative $f'(x)$ vanishes at infinity as $|x|^{\gamma-1-\varepsilon}, \varepsilon>0$, and consider $0<\gamma<1$. Under these assumptions, the chain of equalities is true:

$$\begin{aligned}
\left(\mathcal{D}_{\pm}^{\gamma} f\right)(x) &= \frac{1}{\Gamma(1-\gamma)} \frac{d}{dx} \int_0^{\infty} \frac{f(x \mp \xi)\, d\xi}{\xi^{\gamma}} = \\
&= \frac{1}{\Gamma(1-\gamma)} \int_0^{\infty} \frac{f'(x \mp \xi)\, d\xi}{\xi^{\gamma}} = \\
&= \frac{\gamma}{\Gamma(1-\gamma)} \int_0^{\infty} f'(x \mp t)\, dt \int_t^{\infty} \frac{d\xi}{\xi^{1+\gamma}} = \\
&= \frac{\gamma}{\Gamma(1-\gamma)} \int_0^{\infty} \frac{f(x)-f(x \mp \xi)\, d\xi}{\xi^{1+\gamma}} \overset{def}{=} \left(\mathbf{D}_{\pm}^{\gamma} f\right)(x)
\end{aligned} \tag{33.26}$$

The operators $\left(\mathbf{D}_{+}^{\gamma} f\right)(x)$ and $\left(\mathbf{D}_{-}^{\gamma} f\right)(x)$ in Eq. (33.26) are the *Marchaud fractional derivatives* for an unbounded domain. The advantage of this definition is that the integral converges under more general assumptions for the function, not requiring a good behavior at infinity, i.e., a function growing at infinity as $|x|^{\gamma-\varepsilon}$, with $\varepsilon>0$, has a Marchaud fractional derivative. Therefore, the RL derivative and the Marchaud derivative coincide only for a class of functions. Conditions on the equivalence are reported in [35, pp. 224–229]. The extension of Eq. (33.26) to every positive value of γ reads

$$\left(\mathbf{D}_{\pm}^{\gamma}f\right)(x)=\frac{\{\gamma\}}{\Gamma(1-\{\gamma\})}\int_0^{\infty}\frac{f^{([\gamma])}(x)-f^{([\gamma])}(x\mp\xi)d\xi}{\xi^{1+\{\gamma\}}} \tag{33.27}$$

where $f^{([\gamma])}(x)$ denotes the derivative of order equal to the integer part of γ. The Marchaud fractional derivatives in a finite interval are obtained from Eq. (33.26) by continuing the function $f(x)$ by zero beyond the interval $[a,b]$, that is,

$$f^{*}(x)=\begin{cases} f(x) & x\in[a,b] \\ 0 & x\notin[a,b] \end{cases}$$

obtaining the useful relations

$$\left(\mathbf{D}_{a+}^{\gamma}f^{*}\right)(x)=\left(\hat{\mathbf{D}}_{a+}^{\gamma}f\right)(x)+\frac{f(x)}{\Gamma(1-\gamma)(x-a)^{\gamma}},\quad x\in[a,b] \tag{33.28}$$

$$\left(\mathbf{D}_{b-}^{\gamma}f^{*}\right)(x)=\left(\hat{\mathbf{D}}_{b-}^{\gamma}f\right)(x)+\frac{f(x)}{\Gamma(1-\gamma)(b-x)^{\gamma}},\quad x\in[a,b] \tag{33.29}$$

where $\left(\hat{\mathbf{D}}_{a+}^{\gamma}f\right)(x)$ and $\left(\hat{\mathbf{D}}_{b-}^{\gamma}f\right)(x)$ represent the defined integrals

$$\left(\hat{\mathbf{D}}_{a+}^{\gamma}f\right)(x)=\frac{\gamma}{\Gamma(1-\gamma)}\int_a^x\frac{f(x)-f(\xi)}{(x-\xi)^{(1+\gamma)}}d\xi \tag{33.30}$$

$$\left(\hat{\mathbf{D}}_{b-}^{\gamma}f\right)(x)=\frac{\gamma}{\Gamma(1-\gamma)}\int_x^b\frac{f(x)-f(\xi)}{(\xi-x)^{(1+\gamma)}}d\xi \tag{33.31}$$

The Marchaud definition in Eq. (33.26) can be interpreted as formally obtained from $I_{\pm}^{\gamma}f$ replacing γ with $-\gamma$ and subtracting the function $f(x)$ for convergence's sake [35, pp. 112–116]. In this sense, the Marchaud definition has been proved to be the Hadamard finite parts of Riemann–Liouville fractional integrals [35].

33.3 Fractional model of integral non-local elasticity

Recently, Lazopoulos [23] proposed that the strain energy $W(x)$ for a bar of initial length L under an external force field $f(x)$ can be assumed to be composed of two contributions: (i) a local part of the kind $E\varepsilon^2(x)/2$, where E is the longitudinal modulus and $\varepsilon(x)$ is the strain and; (ii) a contribution of non-local nature

$$-\eta\,\varepsilon(x)\left[\left(\mathcal{D}_{a+}^{\alpha}\varepsilon\right)(x)+\left(\mathcal{D}_{b-}^{\alpha}\varepsilon\right)(x)\right]/2 \tag{33.32}$$

with $0<\alpha<1$, where η is a proportionality constant depending on the material. By performing variation of the total stored energy functional and under the particular constraints $u(a)=u(b)=0$, the constitutive relation of the bar

$$E\varepsilon(x)-\eta\left\{\left(\mathcal{D}_{a+}^{\alpha}\varepsilon\right)(x)+\left(\mathcal{D}_{b-}^{\alpha}\varepsilon\right)(x)\right\}=\sigma(x) \tag{33.33}$$

is reported. After some manipulations, he states that Eq. (33.33) may be converted into the constitutive relation of the form

$$\varepsilon(x)=\frac{\eta}{E}\int_a^b\frac{\varepsilon(\tau)}{(x-\tau)^{\alpha}}d\tau+\sigma(x) \tag{33.34}$$

Apart from some misprinting and a more substantial inaccuracy (in Eq. 16 of [21]) the main idea of Lazopoulos is worthy to be mentioned since he introduces fractional calculus in the field of non-local continuum. Hereinafter, the right way to derive a formula similar to Eq. (33.34) from the strain energy is re-proposed. In this context, let us define the strain energy $W(x)$ with the non-local contribution in the form

$$-\eta\,\varepsilon(x)\left[\left(\mathcal{D}_{a+}^{-\beta}\varepsilon\right)(x)-\left(\mathcal{D}_{b-}^{-\beta}\varepsilon\right)(x)\right]/2 \tag{33.35}$$

with $0<\beta<1$, substantially different from Eq. (33.32). By performing variation of the total stored energy, the constitutive relation takes the form:

$$E\varepsilon(x)-\eta\left\{\left(\mathcal{D}_{a+}^{-\beta}\varepsilon\right)(x)-\left(\mathcal{D}_{b-}^{-\beta}\varepsilon\right)(x)\right\}=\sigma(x) \tag{33.36}$$

From the equivalence

$$\left(\mathcal{D}_{a+}^{-\beta}\varepsilon\right)(x)=\left(I_{a+}^{\beta}\varepsilon\right)(x);\quad \left(\mathcal{D}_{b-}^{-\beta}\varepsilon\right)(x)=-\left(I_{b-}^{\beta}\varepsilon\right)(x) \tag{33.37}$$

Eq. (33.36) may be cast in the form

$$\begin{aligned}\sigma(x) &= E\varepsilon(x)-\eta\left\{\left(I_{a+}^{\beta}\varepsilon\right)(x)+\left(I_{b-}^{\beta}\varepsilon\right)(x)\right\} \\ &= E\varepsilon(x)-\frac{\eta}{\Gamma(\beta)}\int_a^b \frac{\varepsilon(\xi)}{|x-\xi|^{1-\beta}}d\xi = \\ &= E\varepsilon(x)-\eta\int_a^b \varepsilon(\xi)\tilde{g}(x,\xi)d\xi\end{aligned} \tag{33.38}$$

where $\tilde{g}(x,\xi)$ is the inverse power law attenuation function of order $0<(1-\beta)<1$, representing the influence of the strain field at location ξ and the stress field at location x. Equation (33.38), with $a\to-\infty$ and $b\to\infty$ coincides with the constitutive relation of the well-known integral model of non-local elasticity that, in the case of unbounded domain, is

$$\sigma(x)=E\varepsilon(x)-\eta\int_{-\infty}^{\infty}\varepsilon(\xi)\tilde{g}(x,\xi)\,d\xi \tag{33.39}$$

The particular choice of such an attenuation function is very attractive because the parameter β yields large variety of long-range interactions of non-local nature. For comparison's sake with the model presented in this chapter, it is interesting to express Eq. (33.39) in terms of displacements by means of the equilibrium equation $\sigma'(x)=-f(x)$ and with the help of the composition rules of fractional operators reported in the appendix. Equation (33.39) is then recast, after some algebra, in the form

$$E\frac{d^2u(x)}{dx^2}-\eta\left\{\left(\mathcal{D}_+^{2-\beta}u\right)(x)+\left(\mathcal{D}_-^{2-\beta}u\right)(x)\right\}=-f(x) \tag{33.40}$$

and by the introduction of the coefficient $c_\beta=\eta/E$ it becomes

$$\frac{d^2u(x)}{dx^2}-c_\beta\left(\left(\mathcal{D}_+^{2-\beta}u\right)(x)+\left(\mathcal{D}_-^{2-\beta}u\right)(x)\right)=-\frac{f(x)}{E} \tag{33.41}$$

that is, a fractional differential equation involving the left and right Riemann–Liouville fractional derivatives on the infinite axis (Eq. 33.24). As reported in the previous section, for a class of function the RL fractional derivatives coincide with the Marchaud definition, then the latter equation can be expressed also in the equivalent form

$$\frac{d^2u(x)}{dx^2}-c_\beta\left(\left(\mathbf{D}_+^{2-\beta}u\right)(x)+\left(\mathbf{D}_-^{2-\beta}u\right)(x)\right)=-\frac{f(x)}{E}\ ;\ 0\le\beta\le1 \tag{33.42}$$

Despite the formal equivalence of Eqs. (33.41) and (33.42), the Marchaud fractional derivatives appear naturally in a consistent mechanical representation of the non-local bar, as reported in the next sections.

Formulation of the problem for a confined bar in the range $x \in [0, L]$ may be easily obtained replacing the RL fractional integrals $I_{\pm}^{\alpha} f$, valid in the whole axis, with the RL definition on the interval, that is, $I_{0+}^{\alpha} f$ and $I_{L-}^{\alpha} f$ from Eq. (33.3) and Eq. (33.6). By substitution, the governing equation of the elastic problem, in terms of the Marchaud fractional derivatives, assumes the form

$$\frac{d^2u(x)}{dx^2} - c_\beta \left(\left(\mathbf{D}_{0+}^{2-\beta} u \right)(x) + \left(\mathbf{D}_{L-}^{2-\beta} u \right)(x) \right) = -\frac{f(x)}{E}. \tag{33.43}$$

The analysis of boundary value problem described in Eq. (33.43) shows that the non-integral terms involved in Marchaud fractional derivatives on finite support possess divergent nature at the borders. Moreover, there is another fundamental observation arising in Eq. (33.43) involving the presence of supplementary divergent terms also for the first derivative of the displacement function at the borders. This mathematical behavior does not have mechanical explanation at the present time leading to conclude that the requirement of a mathematically and mechanically consistent model is imperative dealing with enriched continuum with cohesive interactions. Some unrealistic effects at the borders of the Eringen model are evidenced also with other types of attenuation functions. In the opinion of the authors the main drawback in the strong non-locality model in Eqs. (33.33, 33.34, 33.35, 33.36, 33.37, and 33.38) is due to the fact that the governing equation is postulated without underlying mechanical model. Long-range interactions will be described in the next section from a different perspective introducing a physical representation.

33.4 Elastic bar with long-range interactions: Unbounded domain

Let us consider an elastic bar with infinite length, as depicted in Fig. 33.1, loaded with external self-equilibrated volume forces denoted $f(x)$ and let us discretize the bar in volume elements $V_j = A\Delta x$ $(j = -\infty, \dots, \infty)$, with A the cross section and Δx the length of the element. Volume element V_j is located at abscissa $x_j = (j-1)\Delta x$ and it is in equilibrium under external loads, contact forces provided by adjacent volume elements, V_{j-1} and V_{j+1}, denoted N_j and N_{j+1}, respectively, and the resultant of long-range actions Q_j applied on V_j by the surrounding non-adjacent elements of the bar (Fig. 33.2).

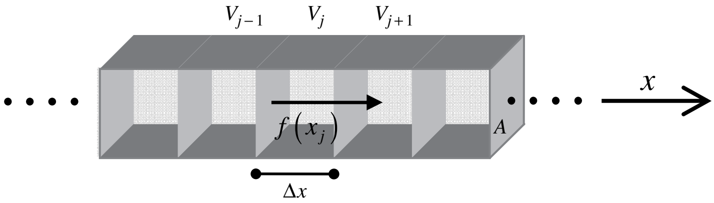


Fig. 33.1. Elastic bar discretized

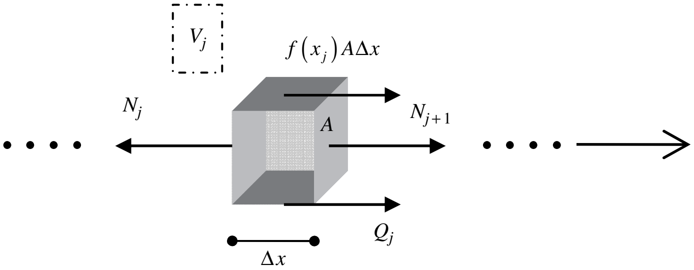


Fig. 33.2. Equilibrium of volume

Under these circumstances the equilibrium equation of volume V_j is provided as

$$\Delta N_j + Q_j = \Delta N_j + \sum_{m=j+1}^{\infty} Q^{(m,j)} - \sum_{m=-\infty}^{j-1} Q^{(m,j)} = -f_j A \, \Delta x \tag{33.44}$$

where $f_j = f(x_j)$, $\Delta N_j = N_{j+1} - N_j$ is the difference between the contact forces N_j and N_{j+1} provided by volume elements V_{j+1}, V_{j-1} and $Q^{(h,j)}$ are the long-range forces that the surrounding volume elements V_h $(h = -m, ..., -2, -1,\ 0,\ 1, 2, ..., m\ (m \to \infty),\ h \neq j-1, j, j+1)$ apply on element V_j as in Fig. 33.3 where only long-range forces have been highlighted. The long-range forces $Q^{(h,j)}$ $(h = -\infty, ..., 0, ..., h \neq j, ..., \infty)$ represent molecular interactions between non-adjacent volume elements and hence they depend on both volume size V_j and V_h of interacting volumes as in applied mechanics problems with interacting axial molecular forces. In the following, long-distance interactions $Q^{(h,j)}$ will be modeled as

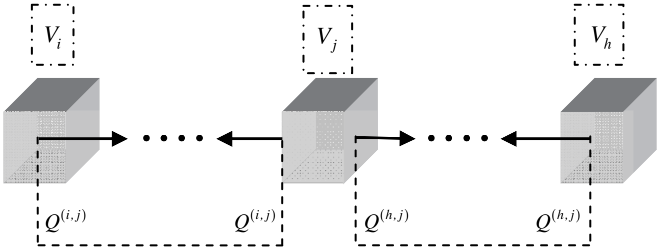


Fig. 33.3. Long-range terms in equilibrium of volume

forces depending on the products of volume elements V_j and V_h as well as the relative displacement $u(x_h)-u(x_j)$ and on a decaying function $g(x_h,x_j)$, that is,

$$Q^{(h,j)} = sign(x_h - x_j)(u(x_h)-u(x_j))g(x_h,x_j)\ V_j\ V_h \tag{33.45}$$

where $sign(x)$ is the well-known signum function defined as

$$sign(x) = \begin{cases} -1 & ;\ x<0 \\ 1 & ;\ x \geq 0 \end{cases} \tag{33.46}$$

The selected decaying function $g(x_h,x_j)$ is a real-valued function expressed as

$$g(x_h,x_j) = \frac{E\ c_\alpha \alpha}{A\ \Gamma(1-\alpha)\left|x_h - x_j\right|^{1+\alpha}};\quad (0 \leq \alpha \leq 1) \tag{33.47}$$

Of course, classical continuum mechanics model, without long-range forces, may be recovered as $\alpha \to 0$. Direct substitution of Eq. (33.47) in the equilibrium equation (Eq. 33.44) yields the equilibrium equation of volume V_j that, under the assumption $V_j = V_h = V_r = A\Delta x$, may be written as

$$\Delta N_j - \frac{E\ c_\alpha \alpha\ A\Delta x}{\Gamma(1-\alpha)}\left[\sum_{h=-\infty}^{j-1}\frac{u(x_j)-u(x_h)}{(x_j-x_h)^{1+\alpha}}\Delta x + \right. \tag{33.48}$$

$$\left. + \sum_{r=j+1}^{\infty}\frac{u(x_j)-u(x_r)}{(x_r-x_j)^{1+\alpha}}\ \Delta x\right] = -f_j A\Delta x$$

Dividing Eq. (33.48) by Δx and taking limit for $\Delta x \to 0$, the differential equilibrium equation obtained is

$$\frac{dN(x)}{dx} - E\ c_\alpha A\left(\left(\mathbf{D}_+^\alpha u\right)(x)+\left(\mathbf{D}_-^\alpha u\right)(x)\right) = -f(x)A \tag{33.49}$$

Equation (33.49) may be recast in terms of conventional stress $\sigma(x) = N(x)/A$ as

$$\frac{d\sigma(x)}{dx} - E\ c_\alpha\left(\left(\mathbf{D}_+^\alpha u\right)(x)+\left(\mathbf{D}_-^\alpha u\right)(x)\right) = -f(x) \tag{33.50}$$

Equation (33.50) is the equilibrium equations of the volume $dV = Adx$ located at abscissa x in which long-range interactions between surrounding non-adjacent volumes have been taken into account.

Assuming linear elastic material, the conventional stress–strain relation may be used:

$$\sigma(x) = E\varepsilon(x) = E\,du/dx \tag{33.51}$$

then, the equilibrium equation in terms of the displacement field for the infinitesimal volume can be written in the form

$$\frac{d^2u(x)}{dx^2} - c_\alpha\left(\left(\mathbf{D}_+^\alpha u\right)(x)+\left(\mathbf{D}_-^\alpha u\right)(x)\right) = -\frac{f(x)}{E} \tag{33.52}$$

Direct comparison of Eq. (33.52) with Eq. (33.42) shows that the two equations are only formally coalescing for unbounded domain since the fractional order of differentiation of Eq. (33.42) is different from that reported in Eq. (33.52). Indeed, both formulations coincide only for $\beta = \alpha = 1$. From these considerations, it appears that postulating a form of the total strain energy, the resulting governing equation is not fully consistent with the proposed mechanical model of long-range forces. Moreover, the mechanical representation of Marchaud fractional derivatives of displacement functions in the fractional integral model of non-local interactions can now be highlighted: they represent the resultant of long-range interactions in the equilibrium of volume $dV = A\,dx$.

Summing up, if we select the attenuation function as in Eq. (33.47) and assume that long-distance interactions $Q^{(h,j)}$ are as reported in Eq. (33.45), then the differential equation in terms of displacements is an ordinary fractional differential equation coalescing with Eq. (33.42) obtained directly postulating the form of total stored energy leading to a fractional Eringen model. The reader could guess that the machinery presented in this section could be avoided by direct introduction in the Eringen model of an opportune attenuation function. At this stage, we may only emphasize that long-range interactive forces exploited in Eq. (33.45) now have a clear mechanical interpretation and this has two consequences: (i) the problem of boundary condition in a finite domain will be introduced in a natural way as it will be shown in the next section; and (ii) the continuous model proposed here has a correspondence with a mechanical discrete model as will be reported later on in the course of the chapter. These two main features remain hidden by direct use of the non-local integral model.

33.5 Analysis of finite domain with long-range interactions

In this section, the problem of finite domain with long-range interactions will be treated with the aid of a mechanical interpretation given in the previous section. The problem is introduced considering a bar of finite length L loaded by external axial force field $f(x)$. The same arguments leading to Eq. (33.48) for the equilibrium equation of volume $V_j = A\Delta x$ with $\Delta x = L/m$ and m the total number of volumes, yield the equation

$$\Delta N_j - \frac{E\, c_\alpha\, \alpha\, A\Delta x}{\Gamma(1-\alpha)}\left[\sum_{h=1}^{j-1}\frac{u(x_j)-u(x_h)}{(x_j-x_h)^{1+\alpha}}\Delta x + \right. \\ \left. + \sum_{h=j+1}^{m+1}\frac{u(x_h)-u(x_j)}{(x_h-x_j)^{1+\alpha}}\Delta x\right] = -f_j A\Delta x \tag{33.53}$$

that represents the analogous of the equilibrium equation reported in the previous section, but with finite number of terms due to the finite extension of the bar. Letting $\Delta x \to 0$, the integrodifferential equilibrium equation may be written as

$$\frac{d^2u(x)}{dx^2} - c_\alpha\left(\left(\hat{\mathbf{D}}_{0+}^\alpha u\right)(x) + \left(\hat{\mathbf{D}}_{L-}^\alpha u\right)(x)\right) = -\frac{f(x)}{E} \tag{33.54}$$

where $\left(\hat{\mathbf{D}}_{0+}^\alpha u\right)(x)$ and $\left(\hat{\mathbf{D}}_{L-}^\alpha u\right)(x)$ are defined in Eqs. (33.30 and 33.31).

Direct comparison of terms in Eq. (33.54) with Eq. (33.52) reveals a substantial difference between the differential equation obtained by direct consideration of the attenuation function in the Eringen model (Eq. (33.52)) and that derived on the mechanical model of long-range forces proposed here. In Eq. (33.54) only the integral part of the Marchaud fractional derivative appears instead of the Marchaud fractional derivative on the finite support. The two equations only coincide for a bar of infinite length. It has to be stressed that in Eq. (33.54) the divergent terms at the borders of the bar domain appearing in the integral non-local model are not present.

Boundary conditions associated to Eq. (33.54) involving kinematic conditions may be imposed for the axial displacements at the restrained locations. If some static boundary condition, say an external force F, is applied at the edge, then we must define the overall resultant stress $\sigma(x)$ at cross section x. To this aim, we observe that the following relation holds:

$$\left(\hat{\mathbf{D}}_{0+}^\alpha u\right)(x) + \left(\hat{\mathbf{D}}_{L-}^\alpha u\right)(x) = \frac{\alpha}{\Gamma(1-\alpha)}\frac{d}{dx}\int_x^L\int_0^x \frac{u(\xi_1)-u(\xi_2)}{|\xi_2-\xi_1|^{1+\alpha}}d\xi_1 d\xi_2 \tag{33.55}$$

and Eq. (33.54) may be recast in terms of the overall stress $\sigma(x)$ as

$$\frac{d}{dx}E\left(\frac{du}{dx} - \frac{c_\alpha \alpha}{\Gamma(1-\alpha)}\int_x^L\int_0^x \frac{u(\xi_1)-u(\xi_2)}{|\xi_2-\xi_1|^{1+\alpha}}d\xi_1 d\xi_2\right) = \frac{d\sigma(x)}{dx} = -f(x) \tag{33.56}$$

with the stress $\sigma(x)$ defined by

$$\sigma(x) = E\left(\frac{du}{dx} - \frac{c_\alpha \alpha}{\Gamma(1-\alpha)}\int_x^L\int_0^x \frac{u(\xi_1)-u(\xi_2)}{|\xi_2-\xi_1|^{1+\alpha}}d\xi_1 d\xi_2\right) \tag{33.57}$$

From Eq. (33.57) we may observe that the overall stress $\sigma(x)$ is the sum of local stress $\sigma_l(x) = E\,du/dx$ and a non-local contribution $\sigma_{nl}(x)$ represented by the second term at the right-hand side of Eq. (33.57). After some algebra, the non-local stress contribution can be written as

$$\sigma_{nl}(x) = -c_\alpha E\Big[\big(I_{0+}^{1-\alpha}u\big)(x) - \big(I_{L-}^{1-\alpha}u\big)(x) + -\frac{1}{\Gamma(1-\alpha)}\left(\int_0^x \frac{u(t)\,dt}{(L-t)^\alpha} - \int_x^L \frac{u(t)\,dt}{(t)^\alpha}\right)\Big] \tag{33.57}$$

Mechanical boundary conditions are easily posed as in classical local mechanics by Eq. (33.57): As it does not show any mathematical inconsistence it suffices to set the applied force at the edges F equal to σA. Summing up, if we postulate that the long-range cohesive forces descend by the Eringen model for an infinite bar and we simply redefine the differential fractional operator on finite support, then two divergent boundary terms appear. Instead, if we assume long-range forces based on physical analysis of the finite bar, the divergent terms disappear but the integral operators $\hat{\mathbf{D}}_{0+}^{\alpha}$ and $\hat{\mathbf{D}}_{L-}^{\alpha}$ are now the integral part of the Marchaud derivative and then the usual rules of fractional calculus do not hold. Now we suppose that the discrete form of Eq. (33.54) (that is expressed in Eq. (33.53)) is not known and we want to use the tools of discretization of fractional calculus. This may be provided resorting to fractional finite differences [36].

In this context, introducing a proper discretization of the bar in m intervals of amplitude $\Delta x = L/m$ and representing the fractional differential operator $D^\alpha = \mathbf{D}_{0+}^{\alpha} + \mathbf{D}_{L-}^{\alpha}$ at the material point $x_j = (j-1)\Delta x$ with $j = 1, 2, \ldots, m+1$ by the difference operator $\Delta_{x_i}^{\alpha}$ is given by

$$\big(D^\alpha s\big)\big(x_j\big) = \Delta_{x_j}^{\alpha} s(x) + O(\Delta x) \tag{33.58}$$

where $O(\Delta x)$ means a quantity of order Δx and the fractional difference operator $\Delta_{x_i}^{\alpha}$ is represented as

$$\Delta_{x_j}^{\alpha} s(x) = \frac{\alpha^{-1}}{\Gamma(1-\alpha)}\left\{\sum_{h=1}^{j-1}\Big[\big(x_{j-h+1}\big)^{-\alpha} - \big(x_{j-h}\big)^{-\alpha}\Big] s(x_h) + \right. \\ \left. + \sum_{r=j+1}^{m}\Big[\big(x_{j-r}\big)^{-\alpha} - \big(x_{j+1-r}\big)^{-\alpha}\Big] s(x_r)\right\} \tag{33.59}$$

Discretizing Eq. (33.54) by operator in Eq. (33.59) and neglecting terms of order Δx, an algebraic fractional difference system in the unknown displacement field $u\big(x_j\big)$ is obtained as

$$\frac{EA}{\Delta x}\Delta^2 u\left(x_j\right)-\frac{E\,c_\alpha A\Delta x}{\Gamma(1-\alpha)}\left\{\sum_{h=1}^{j-1}\left(u\left(x_j\right)-u(x_h)\right)\left(\left(x_{j-h}\right)^{-\alpha}-\left(x_{j+1-h}\right)^{-\alpha}\right)+\right. \tag{33.60}$$

$$\left.-\sum_{r=j+1}^{m}\left(u\left(x_j\right)-u(x_r)\right)\left(\left(x_{r-j}\right)^{-\alpha}-\left(x_{r+1-j}\right)^{-\alpha}\right)\right\}=-F\left(x_j\right)\Delta x$$

holding for $j=1,2,...,m+1$ and with the finite differences $\Delta^2 u\left(x_j\right)=u\left(x_{j+1}\right)-2u\left(x_j\right)+u\left(x_{j-1}\right)$ with $F\left(x_j\right)=f\left(x_j\right)A$.

It must be remarked that the approximation involved in fractional finite differences (Eq. 33.60) requires homogeneous boundary condition for the unknown function. This consideration is worthy to be remarked since the approximation scheme in Eq. (33.59) has been proposed in scientific literature with problems involving Riemann–Liouville fractional derivative defined on finite supports. In this context, divergent behavior at the boundaries is overcome with homogeneous boundary conditions [20, p. 272].

System of m algebraic equations reported in Eq. (33.60) in the unknown displacements $u\left(x_j\right)$ of the grid points used to discretize fractional differential equation may be reported in compact form as

$$\mathbf{K}\,\mathbf{u}=\mathbf{f} \tag{33.61}$$

where displacement and force vectors $\mathbf{u}$ and $\mathbf{f}$, respectively, are vectors collecting nodal displacements and nodal external forces given as

$$\mathbf{u}^T=[u_1 \quad u_2 \quad ... \quad u_m]\ ; \qquad \mathbf{f}^T=[f_1 \quad ... \quad ... \quad f_m]A\,\Delta x \tag{33.62}$$

and the non-local coefficient matrix $\mathbf{K}=\mathbf{K}^l+\mathbf{K}^{nl}$ has been introduced in which contact contributions due to adjacent elements have been considered in the tri-diagonal matrix $\mathbf{K}^l$, collecting elements $K^l=EA/\Delta x$ as

$$\mathbf{K}^l=\begin{bmatrix} K^l & -K^l & ... & ... & 0 \\ -K^l & 2K^l & -K^l & ... & 0 \\ ... & ... & ... & ... & ... \\ ... & ... & -K^l & 2K^l & -K^l \\ 0 & ... & ... & -K^l & K^l \end{bmatrix} \tag{33.63}$$

and non-local interactions have been considered in the symmetric, fully populated, matrix:

$$\mathbf{K}^{nl}=\frac{c_\alpha A E\Delta x}{\Gamma(1-\alpha)} \tag{33.64}$$

$$\begin{bmatrix} K_{11}^{nl} & -\Delta x^{-\alpha} & \left[\Delta x^{-\alpha}-(2\Delta x)^{-\alpha}\right] & \dots & \left[((m-1)\Delta x)^{-\alpha}-(m\Delta x)^{-\alpha}\right] \\ -\Delta x^{-\alpha} & K_{22}^{nl} & -\Delta x^{-\alpha} & \dots & \left[((m-2)\Delta x)^{-\alpha}-((m-1)\Delta x)^{-\alpha}\right] \\ \dots & \dots & \dots & \dots & \dots \\ \left[((m-2)\Delta x)^{-\alpha}-((m-1)\Delta x)^{-\alpha}\right] & \dots & \dots & \dots & \dots \\ \left[((m-1)\Delta x)^{-\alpha}-(m\Delta x)^{-\alpha}\right] & \dots & \dots & -\Delta x^{-\alpha} & K_{mm}^{nl} \end{bmatrix}$$

where elements jh of the matrix $\mathbf{K}^{nl}$ are

$$K_{jh}^{nl}=\frac{\Gamma(1-\alpha)}{c_\alpha \Delta x A E}\int_{x_j}^{x_h} g\left(x_j,\xi\right)d\xi, \quad \text{if } (j\neq h)\,; \qquad K_{jj}^{nl}=-\sum_{\substack{h=1\\ h\neq j}}^{m} K_{jh}^{nl} \tag{33.65}$$

with function $g\left(x_j,\xi\right)$ defined in Eq. (33.47). Displacements at the grid points used to discretize the model are provided by inversion of the stiffness matrix $\mathbf{K}$ accounting for the appropriate kinematic and static boundary conditions at the borders of the solid. Formal equivalence of Eq. (33.61) with the solving equations of elastic problems suggests that an elastic mechanical model may be used to represent mechanics of long-range enriched continuum as will be reported in the next section.

33.6 The mechanical equivalent model of non-local bar

At this point a new insight about the mechanics of the problem at hand may be introduced by considering a discrete spring-point model as reported in Fig. 33.4 only for four points to yield understandable drawing. Contact local forces between adjacent particles have been considered by springs with elastic stiffness $K^l = EA/\Delta x$. Long-distance interactions have been introduced by mechanical connections of non-adjacent particles with linear springs with distance-decaying stiffness as $K_{jh}^{nl}=g\left(x_h,x_j\right)$. Under these circumstances model of Fig. 33.4 may be studied by the classical displacement approach, observing that equilibrium equation for the generic node located at abscissa x_j may be written as

$$\begin{cases} K^l u_1 - K^l u_2 - \sum\limits_{h=2}^{m} g(x_1,x_h)(u_h - u_1) = -F_1 \\ -K^l u_{j-1} + 2K^l u_j - K^l u_{j+1} - \sum\limits_{\substack{h=1\\ h\neq j}}^{m+1} g\left(x_h,x_j\right)\left(u_h - u_j\right) sign(h-j) = -F_j \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } j = 2,\ldots,m-1 \\ K^l u_m - K^l u_{m-1} - \sum\limits_{h=2}^{m} g(x_m,x_h)(u_m - u_h) = -F_m \end{cases} \tag{33.66}$$

First terms in Eqs. (33.66) correspond to contact forces and the sums represent non-local forces applied at material particle located at abscissa x_j by the surrounding particles located at abscissa x_h. The right-hand side of Eqs. (33.66) are related to the body forces applied at material particles.

Equilibrium equations reported in Eq. (33.66) may be rewritten in matrix form similar to Eq. (33.61) introducing the non-local stiffness matrix $\mathbf{K} = \mathbf{K}^l + \mathbf{K}^{nl}$ in which we denote $\mathbf{K}^l$ the local stiffness matrix. Equation (33.63) and the non-local interactions have been incorporated into the symmetric, fully populated, non-local stiffness matrix as

$$\mathbf{K}^{nl} = \begin{bmatrix} K_{11}^{nl} & -g(x_2,x_1) & -g(x_3,x_1) & \ldots & -g(x_m,x_1) \\ -g(x_2,x_1) & K_{22}^{nl} & -g(x_3,x_2) & \ldots & -g(x_3,x_2) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -g(x_m,x_1) & \ldots & \ldots & \ldots & K_{mm}^{nl} \end{bmatrix} \tag{33.67}$$

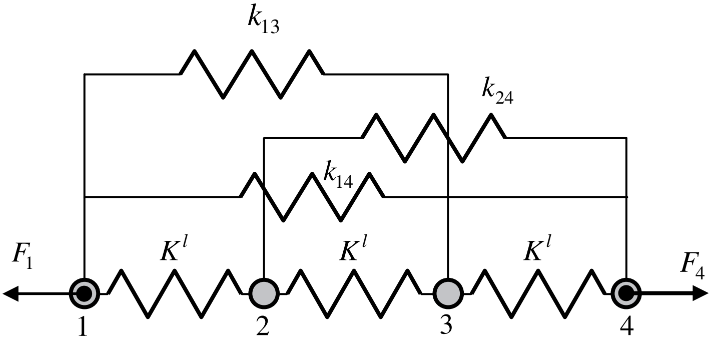


Fig. 33.4. Discrete non-local model with spatially decaying stiffness

where K_{jj}^{nl} has been reported in Eq. (33.65). Moreover, close observation of Eq. (33.67) contrasted with coefficient matrix in Eq. (33.64) shows that when Δx vanishes, the matrix in Eq. (33.67) reverts to the non-local coefficient matrix defined in Eq. (33.64) and obtained with fractional finite differences. This is a very remarkable result enabling us to validate the proposed non-local model of long-range interactions and gives a new perspective in the analysis of enriched continuum.

At this stage, we now have the machinery to represent the mechanical equivalence of the non-integral terms retained in the non-local integral model obtained by the direct use of Eringen model (Eq. 33.43). The divergent boundary terms in Eqs. (33.28 and 33.29) are, in the point-spring non-local model, elastic springs connecting the point to the ground with location-dependent stiffness that reads, at location $x_j = (j-1)\Delta x$ (Fig. 33.5 with four points for clarity of the model):

$$k_j = \frac{E\, c_\alpha \Delta x}{\Gamma(1-\alpha)}\left(\frac{1}{x_j^\alpha} + \frac{1}{\left(L - x_j\right)^\alpha}\right) \tag{33.68}$$

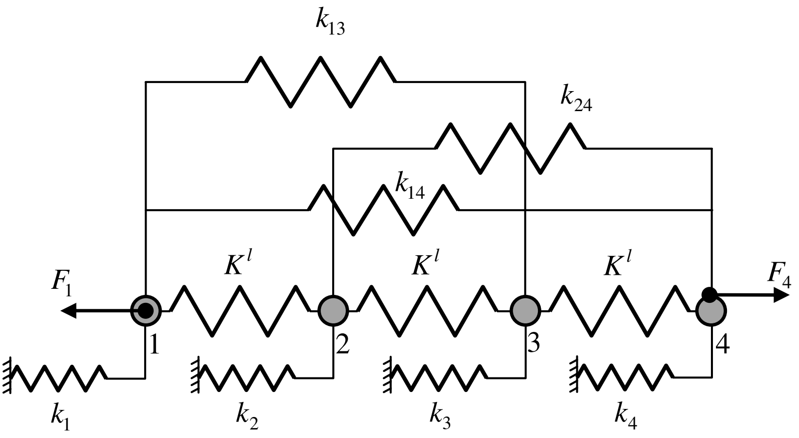


Fig. 33.5. Mechanical equivalence of the Eringen model with point-spring model with additional elastic restraints

Thus, for the model directly derived from the Eringen model the stiffness matrix of the non-local model is provided as $\mathbf{K} = \mathbf{K}^l + \mathbf{K}^{nl} + \mathbf{K}^r$ with the additional, diagonal matrix $\mathbf{K}^r$ of the form

$$\mathbf{K}^r = \begin{bmatrix} k_1 & 0 & \dots & 0 \\ 0 & k_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & k_m \end{bmatrix} \tag{33.69}$$

Because of the form of the additional stiffness in Eq. (33.68) we may conclude that (i) the stiffness of the spring located at the border of the bar is infinitely large corresponding to a fixed support and (ii) the presence of the additional springs is not consistent with the studied bar that is not restrained. These considerations, appearing for the finite bar, are not involved in the analysis of the bar with unbounded domain since in that latter case the stiffness of the additional springs is vanishing everywhere.

As conclusion, Eq. (33.43) directly derived postulating the form of the total energy stored model to be inconsistent from a mechanical point of view because it corresponds to some additional restraints and spring connections that are not present in the mechanical model. It follows that the only way to define the non-local model of a finite bar is that provided in Eq. (33.54) that contains $(\hat{\mathbf{D}}_{0+}^{\alpha}u)(x)$ and $(\hat{\mathbf{D}}_{L-}^{\alpha}u)(x)$ instead of the Marchaud fractional derivatives on finite supports $(\mathbf{D}_{0+}^{\alpha}u)(x)$ and $(\mathbf{D}_{L-}^{\alpha}u)(x)$.

Several numerical investigations conducted by the authors have shown that this concept of the mechanical representation of long-range interactions holds true also for different classes of attenuation functions (Gaussian, Mexican hat, and exponential).

33.7 Dynamics of the non-local fractional model

Dynamic behavior of the non-local model described in the previous section may be formulated accounting for the inertial forces arising during vibrations. To this aim, let us consider first the discrete spring mass model in Fig. 33.4 under the assumption of homogeneous elastic bar with ρ, the mass density of the material and the equilibrium equation of volume V_j, as shown in Fig. 33.6

$$-(\rho A \Delta x)\ddot{u}_j(t) + \Delta N_j^{(l)}(t) + N_j^{(nl)}(t) = 0 \tag{33.70}$$

in which the explicit time-dependence of the state variables terms with $u_j(t) = u(x_j, t)$, M_j the displacement and V_j the mass of the volume, respectively. Inertial forces reported in Eq. (33.70) have been expressed by the time derivative of the axial displacements denoted by $\ddot{u}_j(t) = d^2u_j(t)/dt^2$.

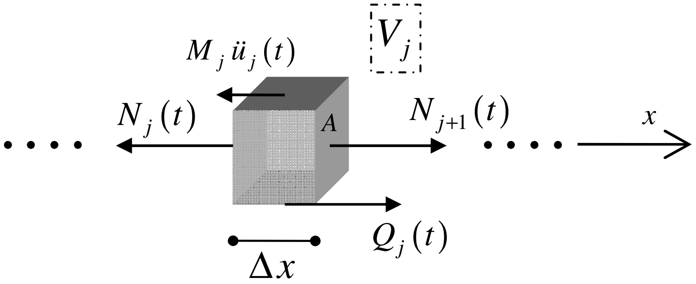


Fig. 33.6. Equilibrium of volume with inertial forces

Substitution of the local and non-local forces, as done in the previous section, leads to the equation of equilibrium that reads

$$-M_j\ddot{u}_j(t) + \Delta N_j(t) + \\ -\frac{Ec_\alpha \alpha A\Delta x}{\Gamma(1-\alpha)}\left[\sum_{h=-m}^{j-1}\frac{u_j(t)-u_h(t)}{(x_j-x_h)^{1+\alpha}}\Delta x + \sum_{h=j+1}^{m}\frac{u_h(t)-u_j(t)}{(x_h-x_j)^{1+\alpha}}\Delta x\right] = 0 \tag{33.71}$$

As Δx vanishes the limit yields a partial fractional differential equation in the axial displacement field $u(t,x)$ as

$$\rho\frac{\partial^2 u(t,x)}{\partial t^2} - E\frac{\partial^2 u(t,x)}{\partial x^2} + Ec_\alpha\left(\left(\hat{\mathbf{D}}_{0+}^{\alpha}u\right)(x) + \left(\hat{\mathbf{D}}_{L-}^{\alpha}u\right)(x)\right) = 0 \tag{33.72}$$

Solution of the differential equation reported in Eq. (33.72) is provided by the fractional finite difference approach already shown in the previous section. Then, a linear system of ordinary differential equations in the time variable involving local and non-local stiffness matrices is obtained in the form

$$\mathbf{M}\ddot{\mathbf{u}}(t) + \left(\mathbf{K}^l + \mathbf{K}^{nl}\right)\mathbf{u}(t) = \mathbf{0} \tag{33.73}$$

where $\mathbf{M}$ is a diagonal matrix whose *j*th diagonal element is $\rho A\Delta x$.

Classical modal analysis can now be easily applied in order to find the frequencies and the vibration mode shapes for the system represented by Eq. (33.72). Assuming the solution in the form

$$\mathbf{u}(t) = \phi_j \, exp\left[\mathrm{i}\,\omega_j\, t\right] \tag{33.74}$$

with ϕ_j and ω_j, respectively, the jth eigenmode and the natural frequency and $\mathrm{i} = \sqrt{-1}$. Substitution of Eq. (33.74) into Eq. (33.73) yields a homogeneous system of algebraic equations involving symmetric and positive definite matrices. In this setting, a non-degenerate solution may be obtained only for values of the parameters ω_j satisfying the secular equation:

$$\det\left[-\omega_j^2\mathbf{M}+\left(\mathbf{K}^{(l)}+\mathbf{K}^{(nl)}\right)\right]=0 \tag{33.75}$$

with $\det[\bullet]$ denoting the determinant of the argument. Some numerical applications reporting the eigenproperties of the non-local model will be described in the next section.

33.8 Numerical applications

In this section, some applications in both static and dynamic setting will be analyzed in order to highlight the most important concepts presented in this chapter. First we want to emphasize, by means of simple examples, the most important characteristic of the three models examined: (i) the proposed model represented by the governing fractional differential equation, Eq. (33.60); (ii) the point-spring model that has been represented in Fig. 33.4 and whose solution in terms of displacements is obtained by means of the stiffness matrix reported in Eq. (33.66); and (iii) the equivalent mechanical model of the integral non-local formulation of Eq. (33.38) schematized in Fig. 33.5. To this aim, in Figs. 33.7 and 33.8 a comparison between the integral non-local model and the proposed representation of cohesive forces has been reported for two bars of different length L. In every example, the bar has been assumed having Young's modulus $E=72$ kN/mm^2, a cross section of area $A=100\,\text{mm}^2$, $\alpha=0.5$, $c_\alpha=0.05$ as the parameters of the decaying non-local law, the external load has been chosen as $F=10\,\text{kN}$ and we used $m=201$.

The bar has been loaded by self-equilibrated forces applied at fixed distance $d=5$ mm from the central cross section with the purpose of representing edge effects. In Figs. 33.7 and 33.8, the axial displacements of the bar with length $L=200\,\text{mm}$ and with $L=10\,\text{mm}$ are reported, respectively.

In both figures, the axial displacements obtained with the proposed model (continuous line) have been contrasted with the discrete point-spring model of non-local interactions (dots), with the non-local integral model (dashed line), and with the classical local model response.

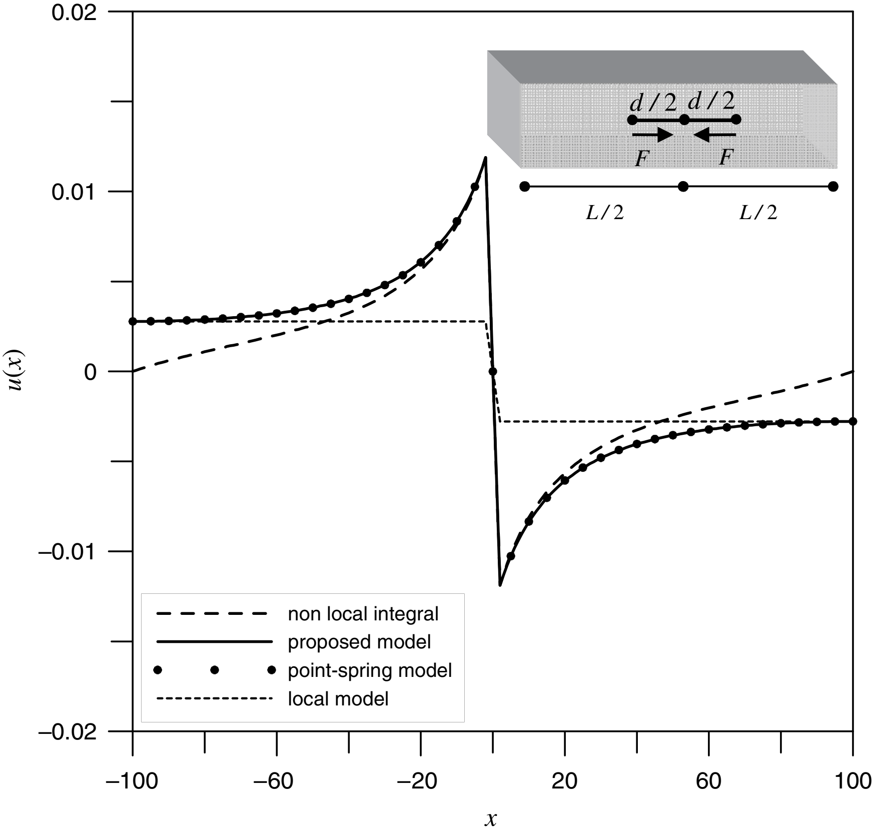


Fig. 33.7. Axial displacements of the proposed fractional model vs the equivalent point-spring model (*dots*), the Eringen model (*dashed*), and the local model for a self-equilibrated bar with $L = 200$ mm

The model proposed matches exactly the point-spring physical model in every case, evidencing therefore its physical consistency. Moreover, the influence of the response of the integral non-local bar on the length of the specimen appears evident. Indeed, for given non-local parameters α and c_α the longer the bar, the fewer the long-range cohesive forces influence the displacements.

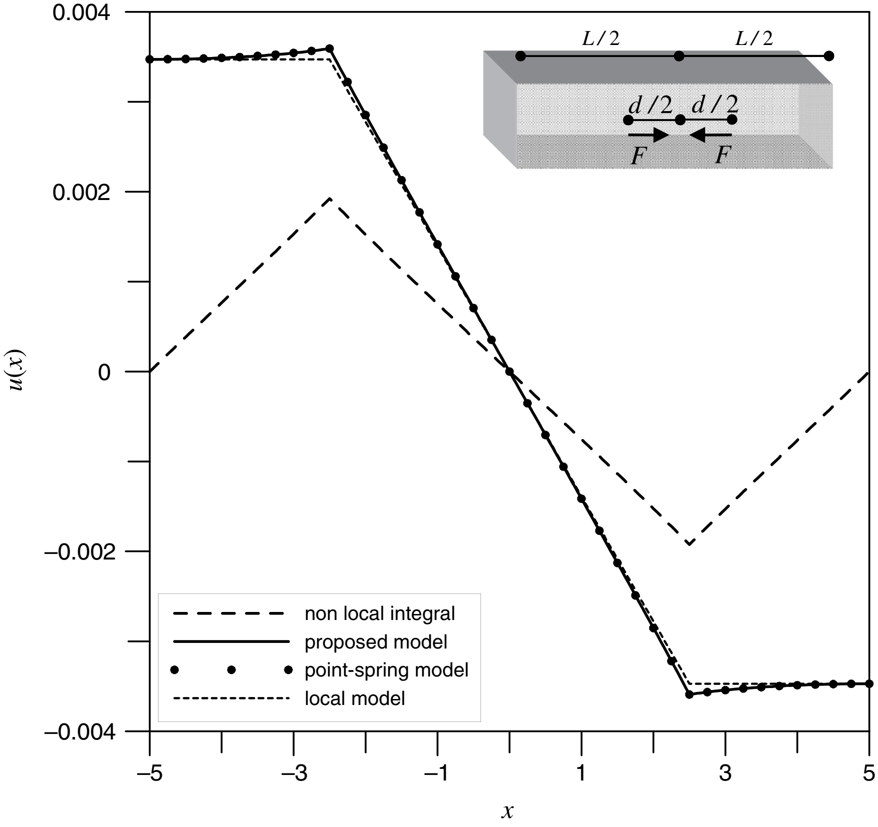


Fig. 33.8. Axial displacements of the proposed fractional model vs the equivalent point-springs model (*dots*), the Eringen model (*dashed*), and the local model for a self-equilibrated bar with $L = 10$ mm

Further, this evidence is true also for the additional constraints in the Eringen model. Indeed, as shown in Fig. 33.7 the non-local integral model yields displacements almost similar to the axial displacement field obtained with the proposed interpretation of long-range forces. Totally different are the displacement in Fig. 33.8 where the specimen is smaller and consequently the effect of the additional constraint is greater. This effect has been discussed in Sect. 33.6 where we showed that the three models coincide only in the case of a bar of infinite length.

Another case of non-local bar handled by the proposed model has been depicted in Fig. 33.9, reporting the axial strains of clamped-free bar under axial forces obtained by the proposed non-local model (continuous line), the equivalent point-spring model (dots), and the local case (dashed line).

The parameter c_α of the non-local model has been selected evaluating the external work done by the applied force in the local case and equating it to the external work done in the non-local case as from Clapeyron work theorem.

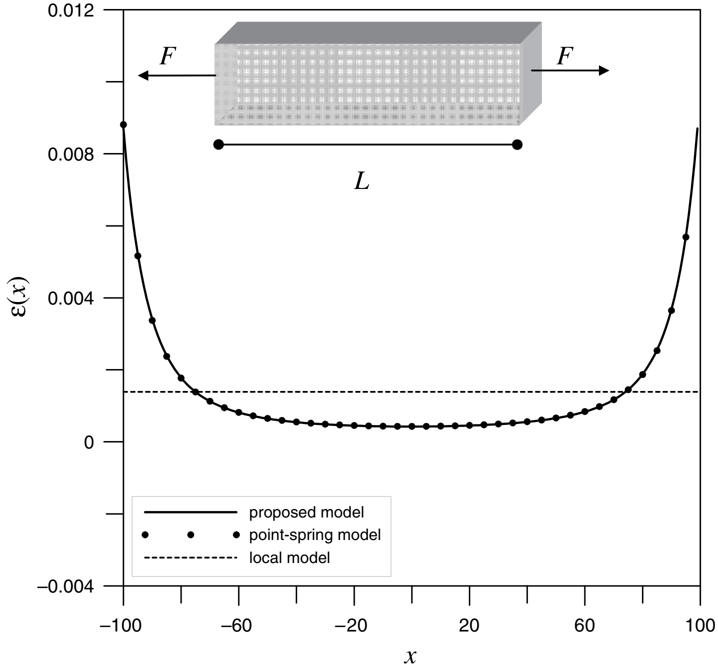


Fig. 33.9. Axial strain of fractional continuum of a free–free bar

The attractiveness of the model proposed relies on the physical solid ground that allows also in passing from the static to the dynamic setting straightforwardly. Mode shapes of the non-local model are in fact found as usual by means of classical modal analysis, evidenced by Eqs. (33.73, 33.74, and 33.75). To this purpose, let us assume that the density of the material is given by $\rho = 2.5\ 10^{-6}\ kg\ \mathrm{mm}/\mathrm{sec}^2$. In Fig. 33.10, the first four eigen-modes are reported contrasting local (dashed), non-local (continuous), and non-local with additional constraints (dot-dashed) eigen-modes.

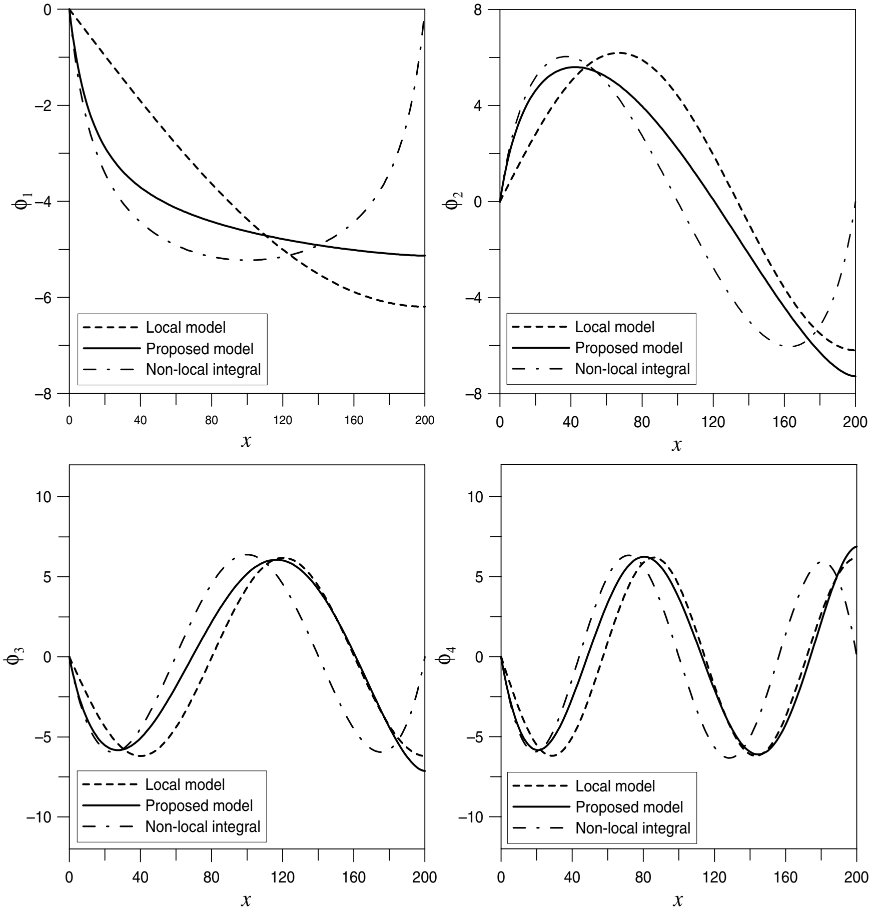


Fig. 33.10. First four eigenmodes of the non-local bar (*continuous line*) contrasted to the eigenmodes of the elastic local bar (*dashed line*)

In Table 33.1, the first six natural frequencies have been reported for the mass-spring non-local model (MS), for the proposed non-local model (PM), for the discrete local (LM) model with $m = 201$ masses, and for the continuum local model (ELM) whose natural frequencies are well known and assume the form $\omega_j = \pi(2j-1)/(2L)\sqrt{E/\rho}$, with $j = 1,2,$.

Table 33.1. Comparison between natural circular frequencies of different models

ω	MS [rad/s]	PM [rad/s]	LM [rad/s]	ELM [rad/s]	Error(%) (LM–PM)	error(%) (ELM–LM)	error(%) (ELM–MS)
ω_1	393.61	393.44	412.27	413.30	4.53	0.25	4.76
ω_2	1305.07	1306.82	1236.79	1239.91	5.52	0.25	5.26
ω_3	1796.96	1795.29	2061.23	2066.52	12.82	0.26	13.04
ω_4	2181.48	2175.92	2885.55	2893.13	24.40	0.26	24.60
ω_5	2520.73	2511.45	3709.69	3719.73	32.05	0.27	32.23
ω_6	2836.97	2824.39	4533.60	4546.34	37.42	0.28	37.60

MS mass-spring model, PM proposed model, LM discrete local model, ELM exact local model

The percentage errors between the investigated models have been reported in the latter columns and some facts appear: (i) first of all the chosen number of masses is consistent because the frequencies of the discrete local model coincide with the corresponding frequencies of the continuum local model; (ii) the frequencies of the proposed model show excellent match with the frequencies of the mass-spring non-local model; and (iii) the error in the evaluation of the natural frequencies between the local and non–local model, reported in the column e% (LM–PM), increases for higher natural frequencies.

Consideration of (iii) is worth to be mentioned since it might explain some differences between the experimentally measured frequencies and the theoretical values often observed in experimental setups.

33.9 Conclusions

In this chapter, the mechanical vibrations of an elastic system with cohesive, long-range interactions have been investigated. The long-range, non-local interactions between the material particles have been included in the investigated model. It has been shown that proper selection of the attenuation function of cohesive interactions as $\left|x_j - x_k\right|^{-(1+\alpha)}$, with $0<\alpha<1$, yields a fractional differential equation of order α. The proposed model is quite different from the well-known Eringen model because non-local interactions have been included on mechanical grounds. It has also been shown that the introduction of the fractional power-law decaying function in the Eringen model yields similar fractional differential equations but with different order of derivation α. This is not a big deal for unbounded domain but it becomes very important for bounded domain as it involves additional restrictions on the function and its first derivatives at the borders. Moreover it has been shown that the fractional model of long-distance interactions coincides with a discrete, particle-spring model that includes connections between non-adjacent particles by linear springs with distance–decaying stiffness. The two models of non-local interactions provide the same mechanical response. These considerations form the basis of the mechanical interpretation of the non-integral terms included in the fractional operators on finite domain. They represent additional external springs with decaying stiffness.

Dynamics of the non-local system has been subsequently investigated including inertial forces in the equilibrium equations leading to a fractional, partial differential equation of order α in the axial displacement of the model. The solution has been obtained "in vacuo" providing the vibration mode shapes and the natural frequencies of the system. Critical comparison between the proposed non-local model and a local system has been reported showing that the mode shapes are similar for the local and non-local models. However, the natural frequencies of the two models are almost the same only for the first two or three frequencies, while they show significant differences at higher eigenfrequencies.

Appendix: Properties of fractional operators

Fractional calculus might involve very cumbersome calculations that can hardly be tackled by hand. In some cases, the composition properties and the Leibniz rule may be helpful and are reported in the following.

A.1 Leibniz rule

In this section, we present the extension of the classical Leibniz rule

$$D^n(f\,g)=\sum_{j=0}^{n}\binom{n}{j}D^{n-j}(f)D^{j}(g) \tag{A.1}$$

to fractional derivatives. It is an important rule because it consents to transform fractional derivatives of many functions in series forms, without performing any burdensome integration. Indeed, given two functions f and g analytic in the interval $[a,b]$, ([35], p. 280), the generalized Leibniz rule is expressed in the form

$$\left(\mathcal{D}_{a+}^{\gamma} f\, g\right)(x)=\sum_{j=0}^{\infty}\binom{\gamma}{j}\left(\mathcal{D}_{a+}^{\gamma-j} f\right)(x)\, g^{(j)}(x) \tag{A.2}$$

where $\binom{\gamma}{j}$ are the binomial coefficients and $g^{(j)}$ is the derivative of integer order j. One must note that the presence of the integer derivative of the function $g(x)$ can be fruitful from the computational perspective. Indeed, the knowledge of the *j*th classical derivative of $g(x)$, combined with the fractional derivative of a constant, for example, $f(x)=1$, gives quite easily a series representation of the fractional derivative of $g(x)$ in the form

$$\left(\mathcal{D}_{0+}^{\gamma} 1\ g\right)(x)=\sum_{j=0}^{\infty}\binom{\gamma}{j}\frac{x^{j-\gamma}}{\Gamma(j-\gamma+1)} g^{(j)}(x) \tag{A.3}$$

with the particular choice $a=0$. The last expression suggests that any fractional derivative, if it exists, has a series representation.

A.2 Compositions rules

Some useful properties dealing with fractional operators will be indicated; for readability's sake, we will report only some simple composition rules referring to [26, 35] for further relations and rigorous proofs. Let us suppose the existences of the integrals and derivatives involved in the following relations, indicated with $\gamma_1>0$ and $\gamma_2>0$, then, the properties

$$I_{a+}^{\gamma_1} I_{a+}^{\gamma_2} f=I_{a+}^{\gamma_1+\gamma_2} f=I_{a+}^{\gamma_2} I_{a+}^{\gamma_1} f \tag{A.4}$$

$$I_{b-}^{\gamma_1} I_{b-}^{\gamma_2} f=I_{b-}^{\gamma_1+\gamma_2} f=I_{b-}^{\gamma_2} I_{b-}^{\gamma_1} f \tag{A.5}$$

on the composition of different order integrals are valid. On the contrary, commutation between fractional integration and differentiation is not straightforward and needs some introductory remarks. In ordinary calculus it is well known that performing first the integral of a function and then a derivative, $\frac{d}{dx}\int_a^x f(x)\,dx=f(x),$ the original function is obtained.

But, if one changes the operations order, i.e, $\int_a^x \frac{d}{dx} f(x)\,dx \neq f(x)$ the result is different because of the presence of a constant. For higher order derivatives and integrals, in the same way $\frac{d^n}{dx^n} I^n_{a+} f = f$ holds, but $I^n_{a+} \frac{d^n}{dx^n} f$ differs from the function by a polynomial of order $n-1$. This simple argument is valid also in the case of fractional operators. Then, the equality

$$\mathcal{D}^{\gamma}_{a+} I^{\gamma}_{a+} f = f(x) \tag{A.6}$$

is always valid and, conversely,

$$I^{\gamma}_{a+} \mathcal{D}^{\gamma}_{a+} f = f(x) \tag{A.7}$$

has been shown [35, p. 43–45] to be true only for those functions having

$$\frac{d^k}{dx^k} I^{1-\{\gamma\}}_{a+} f(a) = 0 \text{ for } k = 1,2,\ldots,[\gamma] \tag{A.8}$$

Also the simultaneous application of integration and differentiation can be simplified in the following way with $f(x) \in Leb_1([a,b])$:

$$\mathcal{D}^{\gamma_2}_{a+} I^{\gamma_1}_{a+} f = I^{\gamma_1 - \gamma_2}_{a+} f, \qquad \mathrm{Re}\gamma_1 > \mathrm{Re}\gamma_2 \tag{A.9}$$

On the contrary, if $f(x)$ does not satisfy Eq. (A8), the relation

$$\left(I^{\gamma}_{a+} \mathcal{D}^{\gamma}_{a+} f\right)(x) = f(x) - \sum_{k=0}^{n-1} \frac{(x-a)^{\gamma-k-1}}{\Gamma(\gamma-k)} \frac{d^{n-k-1}}{dx^{n-k-1}} I^{1-\{\gamma\}}_{a+} f(a) \tag{A.10}$$

holds.

In particular, the composition rules involving classical derivatives read

$$\frac{d^n}{dx^n} \left(\mathcal{D}^{\gamma}_{a+} f\right)(x) = \left(\mathcal{D}^{\gamma+n}_{a+} f\right)(x) \tag{A.11}$$

$$\frac{d^n}{dx^n} \left(\mathcal{D}^{\gamma}_{b-} f\right)(x) = (-1)^n \left(\mathcal{D}^{\gamma+n}_{b-} f\right)(x) \tag{A.12}$$

$$\left(\mathcal{D}^{\gamma}_{a+} \frac{d^n}{dx^n} f(x)\right)(x) = \left(\mathcal{D}^{\gamma+n}_{a+} f\right)(x) - \sum_{j=0}^{n-1} \frac{(x-a)^{j-\gamma-n}}{\Gamma(1+j-\gamma-n)} f^{(j)}(a) \tag{A.13}$$

References

1. Aifantis EC (1994) Gradient effects at macro micro and nano–scales. Journal of the Mechanical Behavior of Materials 5:355–375
2. Aifantis EC (2003) Update on a class of gradient theories. Mechanics of Materials 35:259–280

3. Atanackovic TM (2002) A model for the uniaxial isothermal deformation of viscoelastic body. Acta Mechanica 159:77-86
4. Bažant ZP, Belytschko TB (1984) Continuum theory for strain–softening. Journal of Engineering Mechanics 110:1666–1692
5. Bažant ZP, Jirásek M (2002) Non–local integral formulations of plasticity and damage. Journal of Engineering Mechanics 128:1129–1239
6. Benvenuti E, Borino G, Tralli A (2002) A thermodynamically consistent non–local formulation of damaging materials. European Journal of Mechanics /A Solids 21:535–553
7. Borino G, Failla B, Parrinello F (2003) A Symmetric non–local damage theory. International Journal of Solids and Structures 40:3621–3645
8. Carpinteri A, Chiaia B, Cornetti P (2001) Static–kinematic duality and the principle of virtual work in the mechanics of fractal media. Computer Methods in Applied Mechanics and Engineering 191:3–19
9. Carpinteri A, Chiaia B, Cornetti P (2004) A mesoscopic theory of damage and fracture in heterogeneous materials. Theoretical and Applied Fracture Mechanics 41:43–50
10. Carpinteri A, Chiaia B, Cornetti P (2003) On the mechanics of quasi–brittle materials with a fractal microstructure. Engineering Fracture Mechanics 70:2321–2349
11. Chechkin A, Gonchar V, Klafter J, Metzler R, Tanatarov L (2002) Stationary states of non–linear oscillators driven by Lévy noise. Chemical Physics 284:233–251
12. Cottone G, Di Paola M (2007) On the use of fractional calculus for the probabilistic characterization of random variables. Probabilistic Engineering Mechanics (Submitted)
13. Cottone G, Di Paola M, Pirrotta A (2008) Path integral solution handled by fractional calculus. Journal of Physics: Conference Series, 96: 012007. doi: 10.1088/1742-6596/96/1/012007
14. Di Paola M, Zingales M (2008) Long–range cohesive interactions of non–local continuum faced by fractional calculus. International Journal of Solids and Structures (to appear)
15. Eringen AC, Edelen DGB (1972) On non–local elasticity. International Journal of Engineering Science 10:233–248
16. Fuschi P, Pisano AA (2003) Closed form solution for a non–local elastic bar in tension. International Journal of Solids and Structures 40:13–23
17. Ganghoffer JF, de Borst R (2000) A new framework in non–local mechanics. International Journal of Engineering Science 38:453–486
18. Gonchar V, Tanatarov L, Chechkin A (2002), Stationary solutions of the fractional kinetic equation with a symmetric power–law potential. Theoretical and Mathematical Physics 131(1):582–594
19. Hilfer R (ed) (2000) Applications of fractional calculus in physics. World Scientific Publishing Co
20. Kilbas AA, Srivastava HM, Trujillo JJ (2006) Theory and applications of fractional differential equations. Elsevier, Amsterdam

21. Kroner E (1967) Elasticity theory of materials with long–range cohesive forces. International Journal of Solids and Structures 3:731–742
22. Krumhanls JA (1967) Some considerations of the relations between solid state physics and generalized continuum mechanics. In: Kroner (ed) Mechanics of Generalized Continua. Proc. IUTAM symposium. Springer Verlag, Berlin Heidelberg New York
23. Lazopoulos KA (2006) Non–local continuum mechanics and fractional calculus. Mechanics Research Communication 33:751–757
24. Metzler R, Klafter J (2000) The random walk's guide to anomalous diffusion: a fractional dynamics approach. Physics Reports 339:1–77
25. Metzler R, Klafter J (2004) The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics. Journal of Physics A 37:R161–R208
26. Miller KS, Ross B (1993) An introduction to the fractional calculus and fractional differential equations. John Wiley & Sons, New York
27. Mindlin RD, Eshel NN (1968) On first strain–gradient theories in linear elasticity. International Journal of Solids and Structures 4:109–124
28. Narahari ABN, Hanneken JW, Clarke T (2004) Damping characteristic of a fractional oscillator. Physica A 339:311–319
29. Oldham KB, Spanier J (1974) The fractional calculus. Academic Press, New York
30. Pijaudier–Cabot G, Bažant ZP (1987) Non–local damage theory. Journal of Engineering Mechanics 113:1512–1533
31. Podlubny I (1999) Fractional differential equations. Academic Press, New York
32. Polizzotto C, Borino G (1998) A thermodynamics–based formulation of gradient–dependent plasticity. European Journal of Mechanics A/Solids 17:741–761
33. Polizzotto C (2001) Non–local elasticity and related variational principles. International Journal of Solids and Structures 38:7359–7380
34. Polizzotto C (2003) Gradient elasticity and non standard boundary conditions. International Journal of Solids and Structures 40:7399–7423
35. Samko GS, Kilbas AA, Marichev OI (1993) Fractional integrals and derivatives, Theory and Applications. Gordon and Breach Science Publishers
36. Shkanukov MK (1996) On the convergence of difference schemes for differential equations with a fractional derivative (in Russian). Dokl. Akad. Nauk. 348:746–748
37. West BJ, Bologna M, Grigolini P (2003) Physics of fractal operators. Springer Verlag, Berlin Heidelberg New York

Chapter 34

Fuzzy control for shape memory alloy tendon-actuated robotic structure

N.G. Bizdoaca[1], A. Petrisor[2], E Bîzdoacă[3], I.Diaconu[1]

[1]Department of Mechatronics, University of Craiova, Blvd. Decebal 107, Romania, nicu@robotics.ucv.ro, diaconu@robotics.ucv.ro
[2]Department of Engineering in Electromechanics, Environment and Industrial Informatics, University of Craiova, Blvd. Decebal 107, Romania,apetrisor@em.ucv.ro
[3]Research Center of Mechatronics and Robotics, Blvd. Decebal 107, Romania,elviranicoleta@yahoo.com

Abstract. Shape memory alloy offers an interesting solution, using the shape transformation of the wire/structure in the moment of applying a thermal-type transformation able to offer the martensitic temperature. In order to assure an efficient control of SMA actuator applied to inverted pendulum, a mathematical model and numerical simulation of the resulting model are required. Due to a particular possibility of SMA actuator connection, a modified dynamics for wire or tendon actuation is presented. For an efficient study, a Simulink block set was developed (block for user configurable shape memory alloy material, configurable block for dynamics of single-link robotic structure, block for user configurable wire/tendon actuation). As conventional control possibilities were explored, the fuzzy control structure applied in this chapter. offers an improved response. A more compact SMA actuation is proposed and experimented. The results are commented.

Keywords. Shape memory alloy, Robotics, Conventional control, Fuzzy control, Tendon actuation

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_34,

34.1 Introduction

The shape memory effect was first noted over 50 years ago; it was not until 1962, however, with the discovery of a nickel–titanium shape memory alloy by Buehler, that serious investigations were undertaken to understand the mechanism of the shape memory effect. The shape memory alloys possess the ability to undergo shape change at low temperature and retain this deformation until they are heated, at which point they return to their original shape. The nickel–titanium alloys, used in the present research, generally referred to as Nitinol, have compositions of approximately 50 atomic% Ni/50 atomic% Ti, with small additions of copper, iron, cobalt, or chromium. The alloys are four times the cost of Cu–Zn–Al alloys, but it possesses several advantages as greater ductility, more recoverable motion, excellent corrosion resistance, stable transformation temperatures, high biocompatibility, and the ability to be electrically heated for shape recovery [1].

Shape memory actuators are considered to be low-power actuators and as such compete with solenoids, bimetals, and to some degree motors. It is estimated that shape memory springs can provide over 100 times the work output of thermal bimetals.

The use of shape memory alloy can sometimes simplify a mechanism or device, reducing the overall number of parts, increasing reliability, and therefore reducing associated quality costs. Because of its high resistivity of 80–89 μΩ-cm, nickel–titanium can be self–heated by passing an electrical current through it [2].

The basic rule for electrical actuation is that the temperature of complete transformation to martensite Mf, of the actuator, must be well above the maximum ambient temperature expected.

34.2 Fuzzy logic control

Fuzzy logic is a method of rule-based decision making used for expert systems and process control that emulates the rule-of-thumb thought process used by human beings. The basis of fuzzy logic is fuzzy set theory which was developed by Lotfi Zadeh in the 1960s [3].

Defining a fuzzy controller, process control can be implemented quickly and easily. Many such systems are difficult or impossible to model mathematically, which is required for the design of most traditional control algorithms. In addition, many processes that might or might not be modeled mathematically are too complex or nonlinear to be controlled with traditional strategies. However, if a control strategy can be described qualitatively by an expert, fuzzy logic can be used to define a controller that emulates the heuristic rule-of-thumb strategies of the expert. Therefore, fuzzy logic can be used to control a process that a human can control manually with expertise gained from experience [4]. The linguistic control rules that a human expert can describe in an intuitive and general manner can be directly translated to a rule base for a fuzzy logic controller [5].

A fuzzy controller is composed of the three calculation steps: Fuzzification, fuzzy inference, and defuzzification. The control strategy [6] based on engineering experience with respect to a closed-loop control application is implemented by linguistic rules integrated in the rule base of the controller.

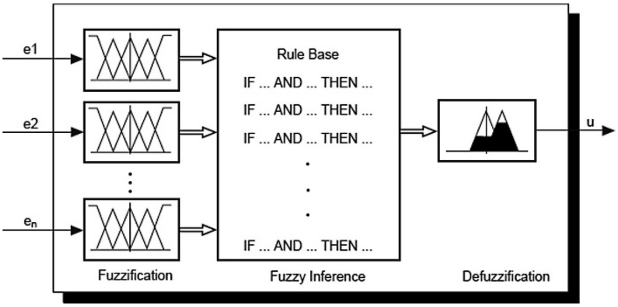


Fig. 34.1. A fuzzy controller structure

Sliding mode control [7, 8] is a type of variable structure control where the dynamics of a nonlinear system is altered via application of a highspeed switching control. This is a state feedback control scheme where the feedback gains are not a continuous function of time. The control scheme involves the following two steps:

- Selection of a hypersurface or a manifold such that the system trajectory exhibits desirable behavior when confined to this manifold
- Finding feedback gains so that the system trajectory intersects and stays on the manifold.

The connection between these two controllers is more than appropriate:

- Both use a rough approximated model of the plant
- Both use a hard forced control
- Both are nonlinear controllers

The manifold border idea of sliding mode controller is related to human experience and expertise for nonlinear plants.

Forcing the outputs of the controller to conduct the system error after a strait trajectory to zero connected with fuzzy approach reveals the direct sliding mode fuzzy controller characteristics [9, 10]. Main disadvantages of this type of controller are the strong outputs dynamics and big value of outputs requirements.

34.3 Applications of shape memory alloy material in robotics

When the alloys and manufacturing techniques improved, so did the experience and results of experimenters. Nitinol received much attention for medical applications, toy industry, teleoperated systems, and robotics, especially autonomous robots.

In 1989 Oaktree Automation Inc., in Alexandria, Virginia, started developing the Fingerspelling Hand, an anthropomorphic robotic device to serve as a tactile communication aid for deaf–blind individuals, particularly those unable to read Braille. The device used a total of one hundred and eight 250 µm Flexinol wires acting in parallel.

The most successful applications of shape memory alloy components usually have all or most of the following characteristics:

- A mechanically simple design
- The shape memory component pops in place and is held by other parts in the assembly
- The shape memory alloy component is in direct contact with a heating/cooling medium
- Friction is minimized and no complex stresses or stress concentrations are present
- A minimum force and motion requirement for the shape memory component
- The shape memory component is isolated from incidental forces with high variation
- The tolerances of all the components realistically interface with the shape memory component

A more efficient support for robotics applications of the SMA wires are SMA springs. The three basic modes in which a shape memory spring can be used are

- constant force;
- constant length;
- simultaneous force and length variation.

34.4 Dynamics of two-link tendon-driven robotic structure

There are many methods for generating the dynamic equations of mechanical system [11]. All methods generate equivalent sets of equations, but different forms of the equations may be better suited for computation or analysis. The Lagrange analysis will be used for the present analysis, a method which relies on the energy properties of mechanical system to compute the equations of motion [12, 13]. We consider that each link is a homogeneous rectangular bar with mass mi and moment of inertia tensor.

$$I_i = \begin{bmatrix} I_{xi} & 0 & 0 \\ 0 & I_{yi} & 0 \\ 0 & 0 & I_{zi} \end{bmatrix} \tag{34.1}$$

Letting $v_i \in R^3$ be the translational velocity of the center of mass for the *i*th link and $\omega_i \in R^3$ be angular velocity, the kinetic energy of the manipulator is

$$T(\theta,\dot{\theta}) = \frac{1}{2} m_1 \|v_1\|^2 + \frac{1}{2} m_1 \omega_1^T I_1 \omega_1 + \frac{1}{2} m_1 \|v_2\|^2 + \frac{1}{2} m_1 \omega_2^T I_2 \omega_2 \tag{34.2}$$

Since the motion of the manipulator is restricted to the xy plane, $\|v_i\|$ is the magnitude of xy velocity of the center of mass and ω_i is a vector in the direction of the y-axis, with $\|\omega_1\| = \dot{\theta}_1$ and $\|\omega_2\| = \dot{\theta}_1 + \dot{\theta}_2$. We solve for kinetic energy, in terms of the generalized coordinates, by using the kinematics of the mechanism.

Using the kinetic energy and Lagrange methods

$$\begin{bmatrix} \alpha + \beta c_2 & \delta + \frac{1}{2}\beta c_2 \\ \delta + \frac{1}{2}\beta c_2 & \delta \end{bmatrix} \begin{bmatrix} \ddot{\theta}_1 \\ \ddot{\theta}_2 \end{bmatrix} + \begin{bmatrix} -\frac{1}{2}\beta s_2 \dot{\theta}_2 & -\frac{1}{2}\beta s_2 \left(\dot{\theta}_2 + \dot{\theta}_1\right) \\ \frac{1}{2}\beta s_2 \dot{\theta}_1 & 0 \end{bmatrix} \bullet \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \end{bmatrix} \tag{34.3}$$

where

$$\alpha = \frac{m_1}{12}\left(l_1^2 + w_1^2\right) + \frac{m_2}{12}\left(l_2^2 + w_2^2\right) + m_1 r_1^2 + m_2\left(l_1^2 + r_2^2\right) \tag{34.4}$$

$$\beta = m_2 l_1 l_2 \tag{34.5}$$

$$\delta = \frac{m_2}{12}\left(l_2^2 + w_2^2\right) + m_2 r_2^2 \tag{34.6}$$

with w_1, w_2, l_1, l_2 the width and the length of link 1 and link 2, respectively.

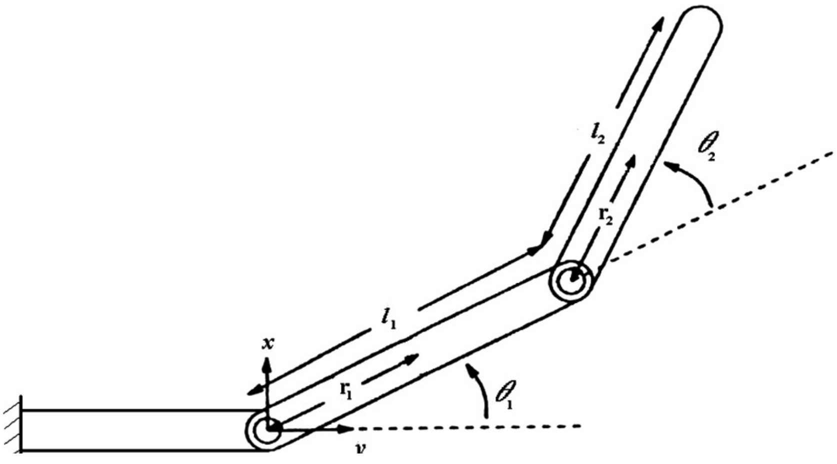


Fig. 34.2. Two-link robotic architecture

34.5 Shape memory actuator structure

Due to the actuation architecture a simple mathematical model can be established. Schematically, the shape memory actuation is given in Fig. 34.3.

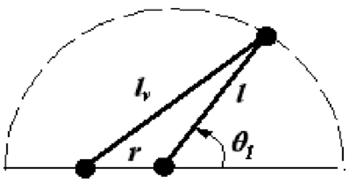


Fig.34 3. Shape memory alloy actuation structure

In Fig. 34.5, l_v is the variable length of shape memory alloy wire, l is the robotic link length between the articulation point and the shape memory alloy wire connection, r is the distance between the second end of the SMA wire (which is a fixed point) and the articulation point of the link (fixed point too) [15].

Using simple mathematical computation the mathematical dependence can be established as

$$\theta_1 = \arccos\left(\frac{l_v^2 - \left(r^2 + l^2\right)}{2lr}\right) \Leftrightarrow \theta_1 = f\left(l_v^2\right) \tag{34.7}$$

The graphic of θ_1 as a function of l_v is given in Fig. 34.4, considering the real domain variation for $\theta_1 \in [0, \pi]$.

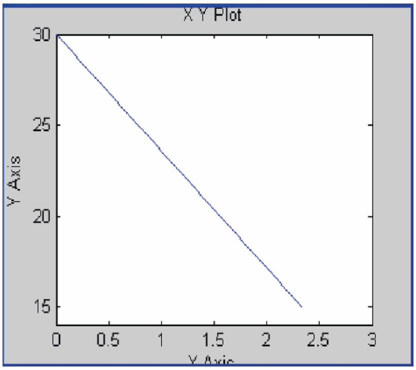


Fig. 34.4. The graphic $l_v = f\left(\theta_1\right)$

As can be easily seen the dependence is linear, and the linearization in modeling can be done successfully. The explanations concern the structural variation of SMA actuator, which are limited superior by l_v and inferior by 0.5 l_v. The mathematical model including the SMA actuation can be developed in two ways. First, it is possible to consider for position control only the length variation of the SMA actuator. This approach is a correct one; the additional torque, provided by the particular properties of SMA, enforces the actuation. The situation corresponds to tendon actuation or wire actuation using the substitution

$$\dot{\theta}_1 = \frac{-2l_v}{lr\sqrt{4 - \left(\frac{l_v^2 - l^2 - r^2}{lr}\right)^2}}\dot{l}_v \tag{34.8}$$

$$\ddot{\theta}_1 = \frac{-2l_v}{lr\sqrt{4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2}}\ddot{l}_v - \frac{2}{lr\sqrt{4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2}}\dot{l}_v^2 - \frac{4l_v^2\left(l_v^2-l^2-r^2\right)}{l^3r^3\sqrt[3]{\left(4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2\right)^2}}\dot{l}_v^2 . \tag{34.9}$$

The mathematical model of the single-link robot with wire actuation is

$$\tau_1 = \left(\frac{m_1w_1^2}{3}\right)\left(\frac{-2l_v}{lr\sqrt{4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2}}\ddot{l}_v - \frac{2}{lr\sqrt{4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2}}\dot{l}_v^2 - \frac{4l_v^2\left(l_v^2-l^2-r^2\right)}{l^3r^3\sqrt[3]{\left(4-\left(\frac{l_v^2-l^2-r^2}{lr}\right)^2\right)^2}}\right) + \frac{gm_1w_1\left(l_v^2-\left(r^2+l^2\right)\right)}{4lr} + b_1\left(\theta_1\right) . \tag{34.10}$$

Analyzing the equilibrium conditions, results that $\tau_1 = b_1\left(\theta_1\right)$ and $l_v^2 = r^2 + l^2$, state which correspond to real case.

The second way makes a simplifying assumption: because the SMA connection with single-link structure can be chosen near to the articulation point, we can assume that the entire SMA torque is directly used for movement. Then the mathematical model can be expressed as

$$\tau_{SMA} = \left(\frac{m_1w_1^2}{3}\right)\ddot{\theta}_1 + \frac{gm_1w_1\cos\left(\theta_1\right)}{2} + b_1\left(\theta_1\right) \tag{34.11}$$

34.6 Fuzzy control applied to shape memory alloy serial-link robotic structure

In order to investigate the SMA robotic structure compartment a Quanser modified platform was used for experiments. The basic control structure uses a configurable PID controller and a Quanser power module unit for energizing the SMA actuators.

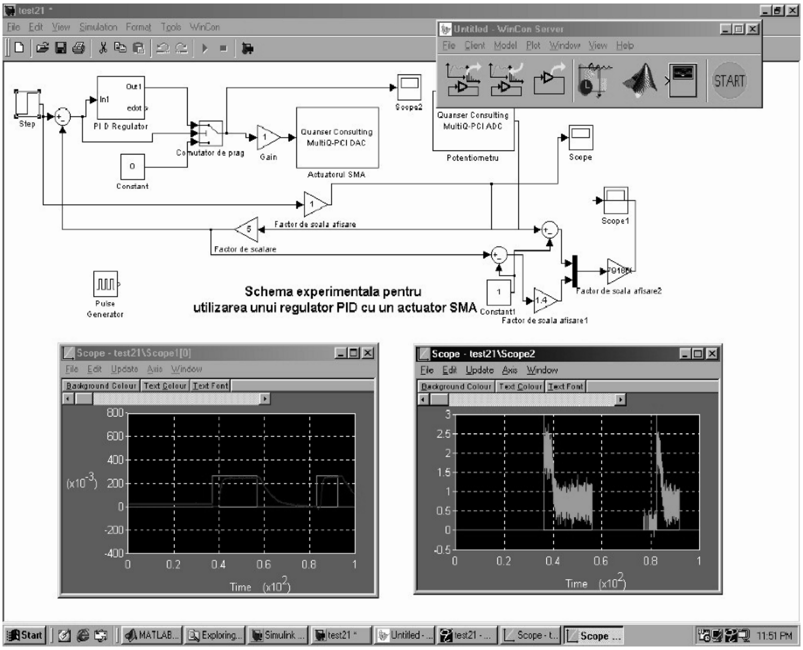

test21 *
File Edit View Simulation Format Tools WinCon
Untitled - WinCon Server
File Client Model Plot Window View Help
START
Step
In1
Out1
edot
PI D Regulator
Constant
0
Comutator de prag
Gain
1
Quanser Consulting MultiQ-PCI DAC
Actuatorul SMA
Scope2
Quanser Consulting MultiQ-PCI ADC
Potentiometru
Scope
Scope1
Factor de scala afisare
5
Factor de scalare
Pulse Generator
Schema experimentala pentru
utilizarea unui regulator PID cu un actuator SMA
Constant1
1.4
Factor de scala afisare1
Factor de scala afisare2
Scope - test21\Scope1[0]
File Edit Update Axis Window
Background Colour Text Colour Text Font
800
600
400
(x10^-3) 200
0
-200
-400
0 0.2 0.4 0.6 0.8 1
Time (x10^2)
Scope - test21\Scope2
File Edit Update Axis Window
Background Colour Text Colour Text Font
3
2.5
2
1.5
1
0.5
0
-0.5
0 0.2 0.4 0.6 0.8 1
Time (x10^2)
Start MATLAB... Exploring... Simulink... test21 * Untitled - ... test21 - ... Scope - t... Scope ...
11:51 PM

Fig. 34.5. Quanser modified platform

PID controller was changed in order to adapt to the particularities of the SMA actuator. A negative command for SMA actuator corresponds to a cooling source.

The actual structure uses for cooling only the ambient temperature.

The best results arise when a PI controller is used. The PI experimented controller parameters are the proportional parameter $K_R = 10$ and the integration parameter $K_I = 0, 05$.

The input step is equivalent with 30° angle base variation and the evolution of this reference is represented with the response of real system in Fig. 34.6.

The control signal variation is presented in Fig. 34.7.

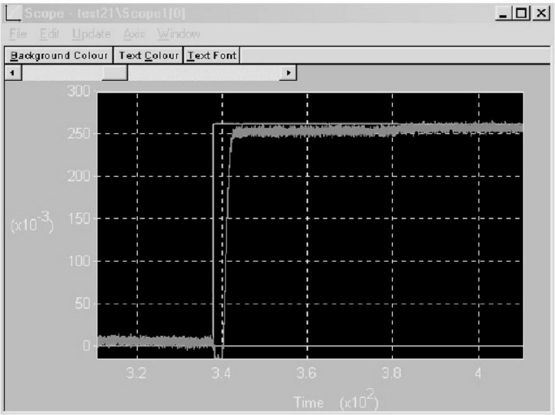


Fig. 34.6. System response, for step input

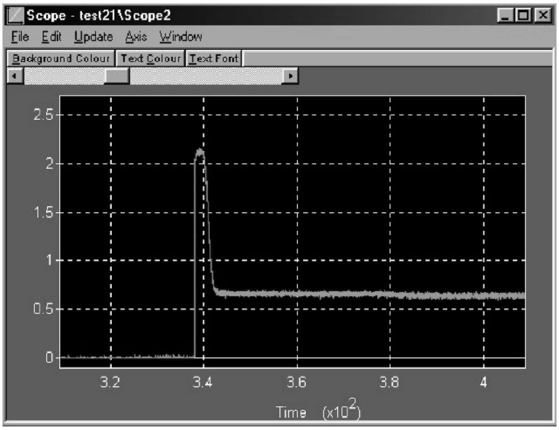


Fig. 34.7. PI controller response, for step input

For negative step, the evolution of the system and the control variable evolution are presented in Figs. 34.8 and Fig. 34.9.

Using PID and PD controller the experiments conduct to less convenient results from the point of view of time response or controller dynamics.

Using heat in order to activate SMA wire, a human operator will increase or decrease the amount of heat in order to assure a desired position to robotic link.

Because of medium temperature influence, a priori, a clear control law, available for all the points of the robotic structure workspace cannot be established. Using the fuzzy theory, a simple and efficient control structure can be implemented.

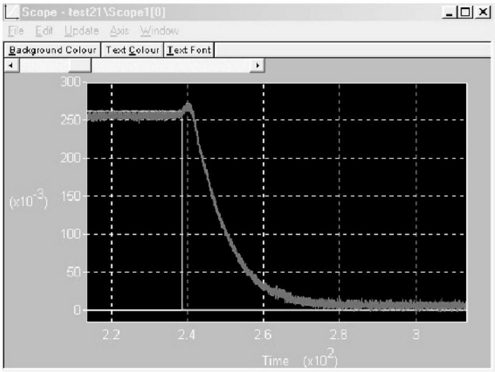


Fig. 34.8. System response, negative step input

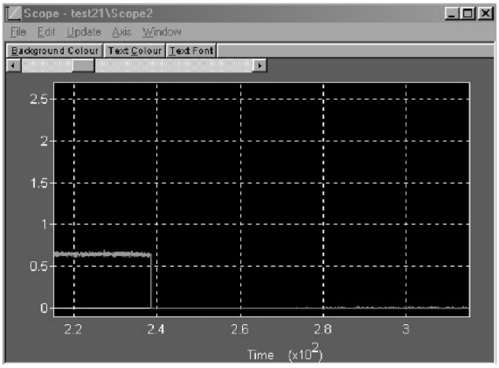


Fig. 34.9. PI controller response, negative step

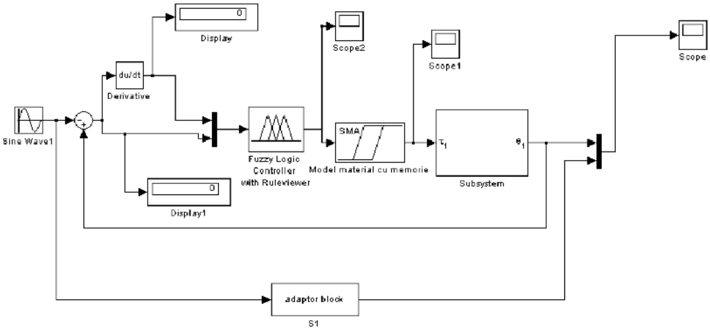


Fig. 34.10. Fuzzy control structure

For an efficient control we propose the following definition for the input and output members:

- input 1 is the first derivate of position error, with three fuzzy members: negative, zero, and positive;

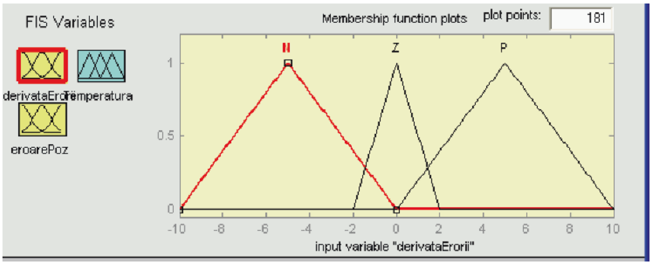


Fig. 34.11. Fuzzy input one member

- input 2 is position error with three fuzzy members: negative, zero, and positive.

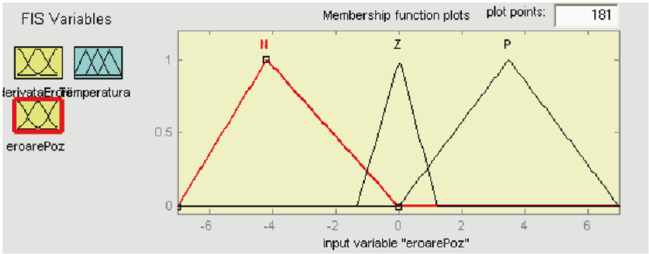


Fig. 34.12. Fuzzy input two members

Output is temperature heating with three fuzzy members: temperature negative (temperature under austenitic start transformation), temperature zero (temperatures between start and final austenitic transformation), and temperature positive (temperature above temperature of final austenitic transformation).

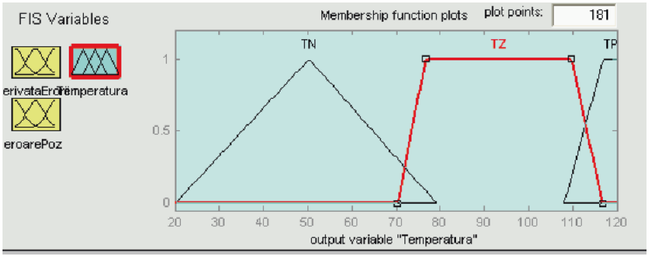


Fig. 34.13. Fuzzy output members

The rules are very simple and are illustrated in the next table.

Table 34.1. The fuzzy rules for proposed controller

	$\dot{e}$	P	Z	N
e				
P		TP	TP	TP
Z		TZ	TZ	TZ
N		TN	TN	TN

The result of the numerical simulation is promising, related to the simplicity of the control structure, for the case of the sinusoidal reference with frequency of 5 rad/s.

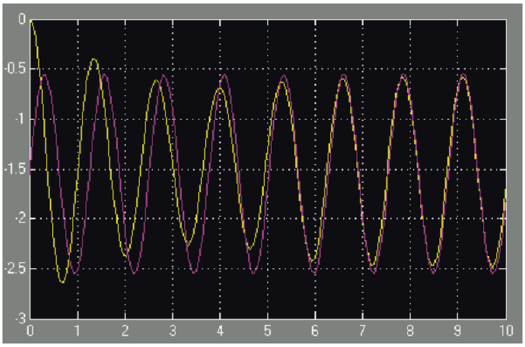


Fig. 34.14. Fuzzy robotic structure output evolution

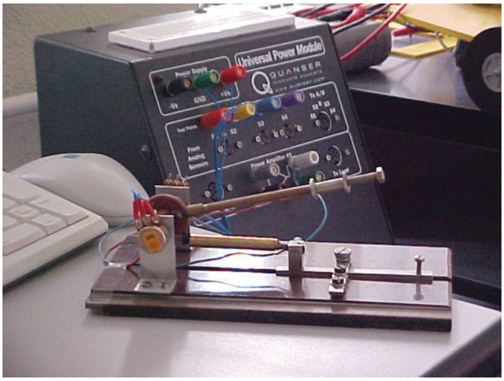


Fig. 34.15. Experimental model for a single-link robotic structure

34.7 Conclusion

The simulations and the mathematical model developed in this chapter offer a background in studying the serial-link robotic control possibilities.

The results respect the real evolution of the structure. In the future, the authors will explore all the control possibilities applied to an extended model [16] and to a real three link model, which for the moment is under construction.

References

1. Funakubo H (1987) Shape Memory Alloys. Gordon and Breach Science Publishers
2. Attanasio M, Faravelli L, Marioni A (1996) Exploiting SMA Bars in Energy Dissipators. Proceedings of the 2nd International Workshop on Structural Control, Hong Kong HKUST, pp 41–50
3. Zadeh L (1965) Fuzzy sets. Information and Control, pp 338–353
4. Ross JT (1995) Fuzzy Logic with Engineering Applications. Mc.Grow Hill, Inc
5. Tao CW (1998) Design of Fuzzy-Learning Fuzzy Controllers. Proceedings of the FUZZ IEEE'98, pp 416–421
6. Soon Yeong Yi (1997) A robust Fuzzy Logic Controller for Robot Manipulators. IEEE Trans. on Systems, Man and Cybernetics, 27(4):706–713
7. Utkin VI (1993) Variable structure systems and sliding model—State of the art assessment. Variable Structure Control for Robotics and Aerospace Applications. New York: K. D. Young Publisher, Elsevier, pp 9–32
8. Utkin VI (1977) Variable structure systems with sliding modes. IEEE Transactions on. Automatic Control. AC-22:212–222
9. Ivanescu M, Stoian V (1995) A Variable Structure Controller for a Tentacle Manipulator. Proceedings of the IEEE on Robotics and Automation. Nagoya, Japan, 3:3155–3160
10. Ivanescu M, Stoian V (1996) A Sequential Distributed Variable Structure Controller for a Tentacle Arm. Proceedings of the IEEE Internation Conference. on Robotics and Automation Minneapolis 4:3701–3706
11. Mason MT (1981) Compliance and Force Control. IEEE Transaction. Systems. Man Cybernetic. 6:418–432
12. Cheng FT, Orin DE (1991) Optimal Force Distribution in Multiple-Chain Robotic Systems. IEEE Transaction. on Systems Man and Cybernetics 21:13–24
13. Cheng FT, Orin DE (1991) Efficient Formulation of the Force Distribution Equations for Simple Closed–Chain Robotic Mechanisms. IEEE Transaction on Systems. Man and Cybernetic. 21:25–32
14. Cheng FT (1995) Control and Simulation for a Closed Chain Dual Redundant Manipulator System. Journal of Robotic Systems, pp 119–133
15. Bîzdoacă NG, Bîzdoacă E, Pană D, Pană C (2003) Shape memory tendon-driven finger. Proceedings of the 14th International Conference on Control Systems and Computer Science. Ed. POLITEHNICA, pp 479–484
16. Ivanescu M (1984) Dynamic Control for a Tentacle Manipulator. Proceedings of Interational Conference, Charlotte, USA

Chapter 35

Implementing delayed file loading functions

Nakhoon Baek[1], Suwan Park[2], Seong Won Ryu[3], Chang Jun Park[3]

[1] School of EECS, Kyungpook National University,
Daegu 702-701, Korea, oceancru@gmail.com
[2] Department of Computer Engineering, Kyungpook National University,
Daegu 702-701, Korea
[3] Digital Content Research Division, Advanced Game Technology Development Center, Electronics and Telecommunications Research Institute,
Daejeon 305-700, Korea

Abstract. Including three-dimensional interactive computer graphics applications, recent application programs often perform the file loading operations to read texture and data files from hard disks. We usually implement these file loading functions as serialized operations, with which the overall programs are suspended to wait for the completion of those file loading operations. In this chapter, we propose a new file loading function, which silently loads the desired file, not to be explicitly recognized by the user. Our delayed file loading functions are designed to satisfy the time constraints on the major work flow. We represent the main ideas and implementation methods for our proposed functionalities. Interface designs for the C++ classes are also presented. We show that the file loading operations are well executed as we originally expected.

Keywords. File loading, Parallel loading, Implementation

35.1 Introduction

File loading operations are frequently used and required by so many programs. Generally, they implement this operation as serialized forms. In

N. Mastorakis, J. Sakellaris (eds.), *Advances in Numerical Methods*,
Lecture Notes in Electrical Engineering 11, DOI 10.1007/978-0-387-76483-2_35,

other words, given a file name, system calls such as open(…) or fopen(…) read the complete contents of the corresponding file and then store the data into a buffer to finally return to the caller function. From the viewpoint of caller functions, it should be wholly suspended to allow the called function entirely perform the file loading operations [1].

In contrast, mainly due to the widespread use of windows systems, event-driven approaches are also required. Since user inputs may be entered in any time, these programs should avoid the suspended situations of the overall programs [2].

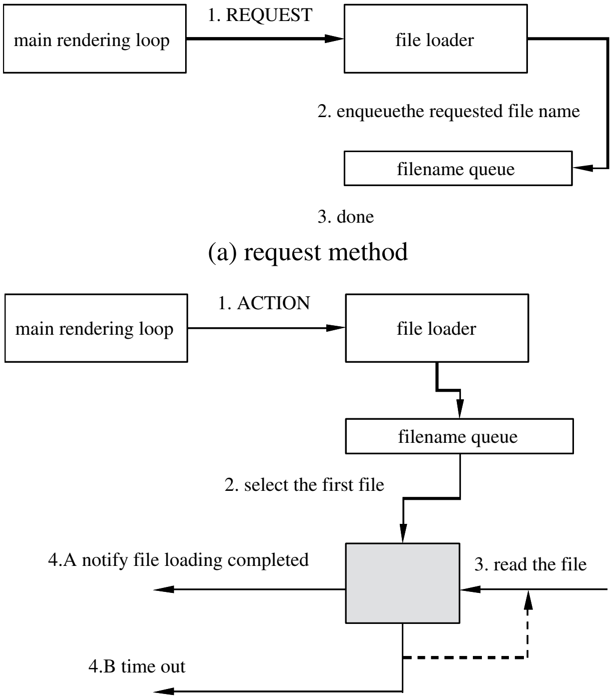
File loading operations are also not an exception. Especially, three dimensional interactive graphics programs usually show sharp drops of their frame rates when the overall programs are suspended. In contrast, recent three-dimensional graphics programs require lots of texture files for high-quality screen outputs. And thus, they need lots of file loading operations.

As a specific example, game application programs require loading of various files such as sound files, map files, and others, in addition to the texture files. In contrast, from the user's point of view, it will be very annoying if the system frequently suspends mainly due to the file loading operations. Thus, in these days, some computer games use more complicated file loading mechanisms, in which recognizing the file loading operations is hard for users [3].

In this chapter, we present a new file loading approach, which performs file loading operations silently, to finally avoid user's explicit recognition of the file loading operations. In Section 35.2, the formal description of the problem is represented. We also show C++ implementation methods in Section 35.3, and our final conclusions are given in Section 35.4.

35.2 Problem formulation

We should consider some constraints on the file loading operations. First, we assume that the file loading module should be executed on a single CPU core, with other time-critical or time-limited operations. Reversely, if the file loading module can be exclusively executed on a single CPU core, a brute-forth approach or immediate execution of requests would be the best strategy. In contrast, according to the overall architecture of real-world applications for PCs or game consoles, the file loading operations have relatively low priorities. Thus, they tend to share a CPU core with other, more important operations.



(a) request method

(b) action method

Fig. 35.1. Method design

As an example, for a typical three-dimensional graphics program compiled for a single-core CPU, the most important constraint is generating 60 frames per second from the rendering module. In this case, the file loading operations would be performed only for the time remaining after the completion of the most important task, the rendering operation. In this configuration, we may meet a failure situation of too-late file loading operations, due to the CPU over-usage of the rendering part. However, typical graphics applications use the texture files mainly to improve visual effects, and file loading failure does not make any critical damage to visual information. Thus, in this chapter, we assume that file loading operations are not time-limited jobs and aim to design a file loading scheme to smoothly load the files without making the overall system suspended.

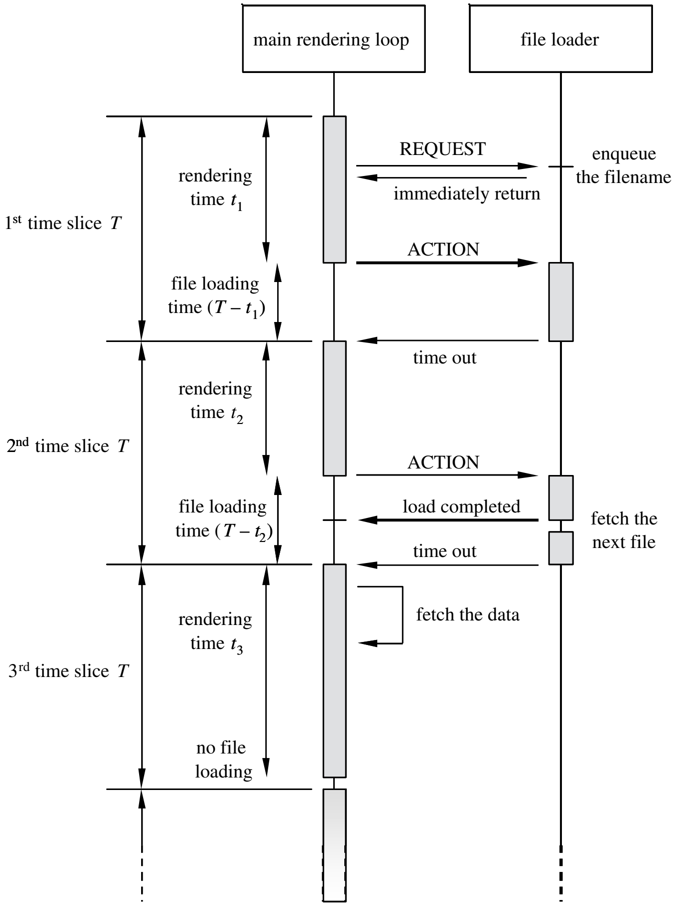


Fig. 35.2. Overall time-transition diagram

For the ease of explanation, we assume that the main rendering part performs all the works except the file loading ones, with some time limits. So, the main rendering part and the file loader do their own works in shifts, while the main rendering part should be repeatedly invoked for every time slice T and thus the execution time period of the file loader may vary depending on that of the main rendering part. Now, the file loader will work only for $(T - t)$ time slices, to satisfy the time constraint of the main rendering part.

As shown in Fig. 35.1(a), the overall scenario starts from instructing the file loader to read some files which would be used for late rendering or calculations, during the execution of some time-limited jobs in the main rendering part. Then, the file loader will enter the file name and its related information into a FIFO queue. From the viewpoint of the file loader, since the current time slice is not its turn, it only records the file name and returns the CPU control to the main rendering loop immediately.

After the rendering job of the main rendering part, there is the remaining time ($T - t$) until the next invoke of the rendering part. Now, as shown in Fig. 35.1(b), the main rendering part passes over the CPU control to the file loader, for the remaining time ($T - t$). The file loader fetches a file name from the FIFO queue and starts to read its content. To keep its own time limit ($T - t$), the file loader reads the file in a segment-by-segment manner, rather than reading the whole content in a single action. When the time limit ($T - t$) is over before completely reading the file content, the file loader stops its operations and passes over the CPU control to the main rendering part, to finally keep the time constraint of the main rendering part.

In the next time slice, the file loader resumes the suspended file reading operation. When a file reading operation is completed, the file loader reports the end of the file loading operation and its resulting data to the main rendering loop. It fetches the next file name and starts the next file loading operation. The overall time-transition diagram is shown in Fig. 35.2.

35.3 Class implementation

We have implemented the file loading schemes explained in Section 35.2 as a C++ class. Major operations are implemented into two methods, as shown in the following:

- **void** request(**char*** *filename*, **bool*** *completed*, **void*** *buf*);
  It requests the loading of the file whose name is *filename*. The parameter *completed* is the flag to indicate the status of the file loading operation. This function resets the *completed* to be false, while it becomes true after loading the file contents to the buffer area *buf*, by the action(…) method shown below. Then, the caller function checks the *completed* flag and uses the data in the *buf* area, when the *completed* flag is true.
- **void** action(**float** *remained*);
  Use the remaining time slice ($T - t$) for the file loading operation. First, it registers an ALARM signal after the remaining time and does its file loading operation. When the file loading is completed before the ALARM signal, the *completed* flag is set to true and the data loaded into the buffer *buf*. For the time-out situations, the current operation is suspended and it returns to the caller function.

Using the above interfaces, we implemented the file loader and its corresponding rendering part. The testing result shows that the delayed file loading works well, as our original goals.

35.4 Conclusion

Many application programs frequently use the file loading operations, and in most cases, they are implemented as serialized operations. In contrast, they need to perform the file loading operations in a parallelized manner. In this chapter, we proposed a new file loading approach whose operations are not easy to recognize explicitly, by performing the file loading operations while satisfying the time constraints on the major work flow. Basic ideas and their corresponding implementations are represented. We also show the interface designs for the C++ classes. We found that the file loading module works well as its original design. In the future, more improvements are needed for more efficient and intuitive implementations.

Acknowledgment. This work was supported by the IT R&D program of MIC/IITA [2007-S001-01, Development of Multi-Core CPU & MPU-Based Cross-Platform Game Technology].

References

1. Glass G, Ables K (2003) Unix for Programmers and Users. 3rd Ed. Pearson/Prentice-Hall, New Jersey
2. Petzold C (2002) Programming Windows. 5th Edition. Microsoft press, Washington
3. Bilas S (2003) The continuous world of dungeon siege. Game Developer's Conference

Author Index

Subject Index

Continued from page ii

Intelligentized Methodology for Arc Welding Dynamical Processes: Visual
Information Acquiring, Knowledge Modeling and Intelligent Control
Chen, Shan-Ben, Wu, Jing
978-3-540-85641-2

Proceedings of the European Computing Conference: Volume 2
Mastorakis, Nikos, Mladenov, Valeri, Kontargyri, Vassiliki T. (Eds.)
978-0-387-84818-1

Proceedings of the European Computing Conference: Volume 1
Mastorakis, Nikos, Mladenov, Valeri, Kontargyri, Vassiliki T. (Eds.)
978-0-387-84813-6

Electronics System Design Techniques for Safety Critical Applications
Sterpone, Luca
978-1-4020-8978-7

Data Mining and Applications in Genomics
Ao, Sio-Iong
978-1-4020-8974-9, Vol. 25

Informatics in Control, Automation and Robotics: Selected Papers from the International Conference on Informatics in Control, Automation and Robotics 2007
Filipe, J.B.; Ferrier, Jean-Louis; Andrade-Cetto, Juan (Eds.)
978-3-540-85639-9, Vol. 24

Digital Terrestrial Broadcasting Networks
Beutler, Roland
ISBN 978-0-387-09634-6, Vol. 23

Logic Synthesis for Compositional Microprogram Control Units
Barkalov, Alexander, Titarenko, Larysa
ISBN: 978-3-540-69283-6, Vol. 22

Digital Terrestrial Broadcasting Networks
Beutler, Roland
ISBN 978-0-387-09634-6, Vol. 23

Logic Synthesis for Compositional Microprogram Control Units
Barkalov, Alexander, Titarenko, Larysa
ISBN: 978-3-540-69283-6, Vol. 22

Sensors: Advancements in Modeling, Design Issues, Fabrication and Practical Applications
Mukhopadhyay, Subhas Chandra; Huang, Yueh-Min (Eds.)
ISBN: 978-3-540-69030-6