Sio-Iong Ao
Len Gelman
Haeng Kon Kim *Editors*

Transactions on Engineering Technologies

World Congress on Engineering 2019

Springer

Transactions on Engineering Technologies

Sio-Iong Ao · Len Gelman · Haeng Kon Kim
Editors

Transactions on Engineering Technologies

World Congress on Engineering 2019

Springer

Editors

Sio-Iong Ao
International Association of Engineers
IAENG Sectretariat
Hong Kong, Hong Kong

Len Gelman
University of Huddersfield
Huddersfield, UK

Haeng Kon Kim
School of Software Convergence
Daegu Catholic University
Daegu, Korea (Republic of)

ISBN 978-981-15-8272-1 ISBN 978-981-15-8273-8 (eBook)
https://doi.org/10.1007/978-981-15-8273-8

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

A large international conference on Advances in Engineering Technologies and Physical Science was held in London, UK, July 3–5, 2019, under the World Congress on Engineering 2019 (WCE 2019). The WCE 2019 is organized by the International Association of Engineers (IAENG); the Congress details are available at: http://www.iaeng.org/WCE2019. IAENG is a non-profit international association for engineers and computer scientists, which was founded originally in 1968. The World Congress on Engineering serves as good platforms for the engineering community to meet with each other and to exchange ideas. The conferences have also struck a balance between theoretical and application development. The conference committees have been formed with over 300 committee members who are mainly research center heads, faculty deans, department heads, professors, and research scientists from over 30 countries. The congress is truly global international event with a high level of participation from many countries. The response to the Congress has been excellent. There have been more than 300 manuscript submissions for the WCE 2019. All submitted papers have gone through the peer-review process, and the overall acceptance rate is 50.93%.

This volume contains 18 revised and extended research articles written by prominent researchers participating in the conference. Topics covered include mechanical engineering, engineering mathematics, computer science, electrical engineering, and industrial applications. The book offers the state of the art of tremendous advances in engineering technologies and physical science and applications, and also serves as an excellent reference work for researchers and graduate students working on engineering technologies and physical science and applications.

Hong Kong — Sio-Iong Ao
Huddersfield, UK — Len Gelman
Daegu, Korea (Republic of) — Haeng Kon Kim

Contents

About the Editors

Dr. Sio-Iong Ao finished his doctoral research at The University of Hong Kong and postdoctoral researches at the University of Oxford and Harvard University and is a former Visiting Professor of Cranfield University, UK, and University of Wyoming, USA.

Prof. Len Gelman, Ph.D. Dr. of Sciences (Habilitation) joined Huddersfield University as Professor, Chair in Signal Processing/Condition Monitoring and Director of Centre for Efficiency and Performance Engineering, in 2017, from Cranfield University, where he worked as Professor and Chair in Vibro-Acoustical Monitoring since 2002.

Len developed novel condition monitoring technologies for aircraft engines, gearboxes, bearings, turbines, and centrifugal compressors.

Len published more than 250 publications, 17 patents and is co-editor of 11 Springer books.

He is Fellow of: BINDT, International Association of Engineers and Institution of Diagnostic Engineers, Executive Director, International Society for Condition Monitoring, Honorary Technical Editor, International Journal of Condition Monitoring, Editor-in-Chief, International Journal of Engineering Sciences (SCMR), Chair, annual International Condition Monitoring Conferences, Honorary Co-Chair, annual World Congresses of Engineering, Co-Chair, International Conference COMADEM 2019 and Chair, International Scientific Committee of Third World Congress, and Condition Monitoring.

He was General Chair, First World Congress, Condition Monitoring, Chair, Second World Congress, Engineering Asset Management and Chair, International Committee of Second World Congress, Condition Monitoring.

Len is Chair of International CM Groups of ICNDT and EFNDT and Member of ISO Technical Committee, Condition Monitoring.

Len made 42 plenary keynotes at major international conferences. He was Visiting Professor at ten universities abroad.

Prof. Haeng Kon Kim is Vice President of Research and Information, Dean of engineering college, and Professor in the Department of Computer Engineering Catholic University of Daegu, Korea. He has been a research staff member in Bell Labs and NASA Center in USA. Professor Kim is Chief Editor of KIPS SE-Sig journal and Korea Multimedia Society, an editorial board of Korea Information Science Society (KISS) and a steering committee of Korea Information Processing Society (KIPS).

Numerical Investigation of an Unmanned Aircraft Vehicle (UAV) Using Fluid-Structure Interaction

Kevin Marangi and Salim Mohamed Salim

Abstract This study employed Fluid-Structure Interaction (FSI), which is the coupling of Computational Fluid Dynamics (CFD) with Finite Element Analysis (FEA), to investigate the structural consequences of a wind gust on an Unmanned Aircraft Vehicle (UAV). The wind gust is modelled as a sudden increase to 23 ms^{-1} in airspeed when the UAV is initially cruising at a velocity of 13 ms^{-1}. In the first step, CFD simulations were carried out using ANSYS FLUENT, and validated against XFLR5 (an open-source software based on Massachusetts Institute of Technology (MIT)'s low Reynolds number CFD program, XFOIL). A steep increase in aerodynamic loads is observed as a result of the wind gust. The values jumped to 244 N for lift and 13.2 N for drag compared to 77.2 and 4.34 N during normal cruise flight conditions. In the next stage, the CFD-obtained pressure fields were exported to ANSYS MECHANICAL to run a structural analysis of the wings' response to the induced aerodynamic load. A slender component connecting the back-wing's outer shell and spar, experienced the largest maximum stress of 75.0 MPa, which amounts to a threefold increase from 23.8 MPa during normal flight conditions. In the final step, the FEA numerical results are analytically calculated to determine the structural response of the wing-fuselage connectors. The entire investigation concludes that, although larger aerodynamic loads, and consequently larger stresses are generated due to an increase in wind speed (mimicking a sudden wind gust), the UAV's structural integrity remains intact.

Keywords Aerodynamic loads · CFD · Fluid-structure interaction · Turbulence · UAV · Composite materials · Canard wing configuration

K. Marangi (✉)
Department of Mechanical Engineering, School of Engineering, EPFL, Route Cantonale, 1015 Lausanne, Switzerland
e-mail: marangikevin1@gmail.com

S. M. Salim
College of Engineering, Swansea University, Swansea SA1 8EN, UK
e-mail: s.m.salim@swansea.ac.uk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_1

1 Introduction

UAVs have become central players in an increasing number of civilian applications despite originally being developed for military use. The Institution of Mechanical Engineers (IMechE) has introduced a competition that challenges university students to design and build a UAV, which could be deployed to assist during disasters. University of Dundee's engineering student team*, Haggis Aerospace*, participated in this competition and designed the UAV in Fig. 1. This UAV featured a *lifting canard wing configuration*, which is a non-traditional aircraft wing configuration. In this arrangement, the smaller flying surfaces, called *canard wings*, are located before the main wing in order to contribute to the total lift generated. It is worth noting that in the conventional wing setup, the larger wing is in front of the smaller flying surface, which is generally used for aircraft control rather than for lift generation.

The student team employed XFL5 software as a CFD tool for designing their UAV, as XFLR5 is based on MIT's famous XFOIL program and is also open source. Due to the additional manner of generating lifting forces in this wing arrangement, it is necessary to investigate that whether the semi-empirical XFLR5 software and the more traditional CFD software, like ANSYS Fluent, can both predict a similar aerodynamic performance of the UAV.

Furthermore, XFLR5 is known to provide satisfactory preliminary simulations (discussed below), but unfortunately it is limited to fluid simulations, which means that although the aerodynamic performance of a UAV can be estimated, it is not possible to predict the UAV's structural response from certain flight conditions. Because of this, it is also vital to investigate the structural integrity of both wings under two types of typical flight conditions: comparing airspeed of 13 m s^{-1} (*normal*

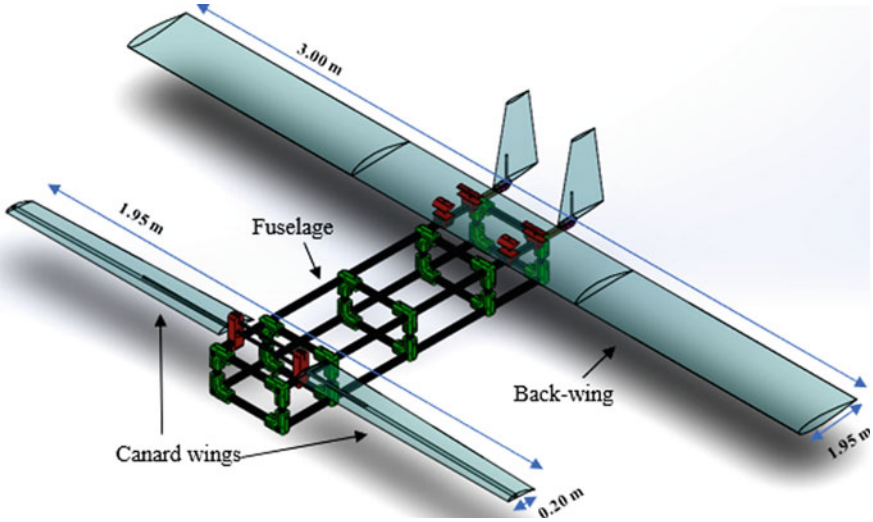


Fig. 1 Present study's Computer-Aided Design (CAD) model

Table 1 Material used in present study

Component	Material
Wings' materials[a]	Woven and unidirectional Carbon Fibre Reinforced Plastic (CFRP); woven Glass-Fibre reinforced Plastic (GFP); Expanded Polypropylene (EPP) foam
Wing-fuselage connector's materials[a]	PETG (3D printed); aluminum alloy bolt

[a] Materials used in the wings and connector are obtained from the student team (Haggis Aerospace) in charge of the UAV [1], but data owned by the author K. Marangi

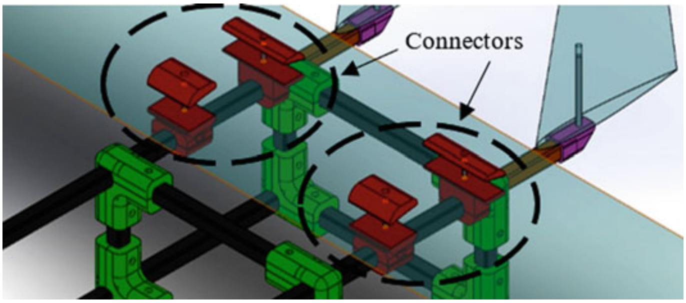


Fig. 2 Close-up view of the components connecting the back-wing and the fuselage

flight condition) to airspeed of 23 m s^{-1} (*wind gust* flight condition). Predicting the structural consequences of a wind gust is beneficial as it would evaluate the current structural performance of the UAV, which would prevent critical damage during flight due to unforeseen airspeed changes. A one-way FSI is employed to carry out this investigation, which is done by coupling CFD and FEA. Table 1 lists the main materials that will be considered in this study. The majority of the components were manufactured using composite materials, except from the back-wing's wing-fuselage connectors using Polyethylene Terephthalate Glycol-modified (PETG) and aluminium nutted bolts to provide clamping power (Fig. 2).

In the present study, ANSYS FLUENT is used to first simulate the airflow around the wings of the UAV for the different wind speeds, providing a large amount of information about the corresponding airflows, such as pressure fields which are flow necessary properties for a FSI analysis.

These pressure fields were then transferred to ANSYS MECHANICAL and mapped onto the structural model of the UAV wings to investigate the resulting stresses. For this stage, the wings are modelled as a 3D cantilever with varying loading, as well as with a fixed end at the wing root. It was also necessary to account for the fact that the wings are manufactured using several materials.

Lastly, in order to analytically determine the stresses on the wing-fuselage connectors, the previously numerically calculated wing stresses obtained from the previous stages are implemented into the analytical problem. It was possible to simplify this connector problem to a simple tensile stress problem thanks to several assumptions

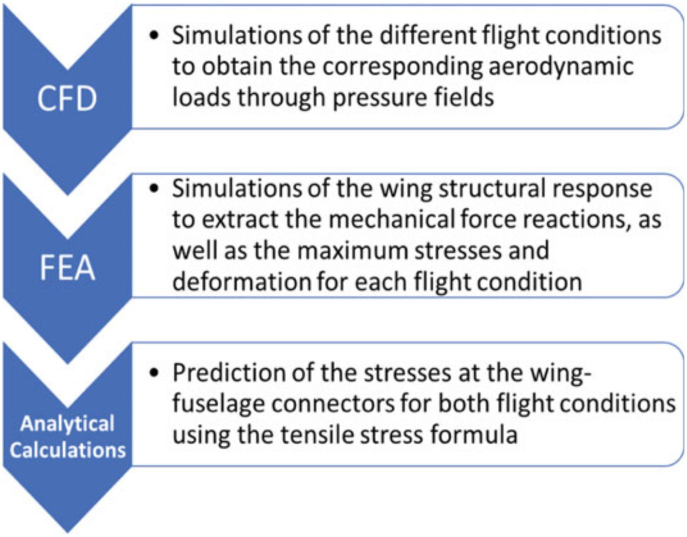

CFD
• Simulations of the different flight conditions to obtain the corresponding aerodynamic loads through pressure fields
FEA
• Simulations of the wing structural response to extract the mechanical force reactions, as well as the maximum stresses and deformation for each flight condition
Analytical Calculations
• Prediction of the stresses at the wing-fuselage connectors for both flight conditions using the tensile stress formula

Fig. 3 Present study's workflow

which are going to be presented in the Methodology section. Figure 3 summarises the workflow for the present study, which also builds on from previous work [2].

UAVs generally operate at lower Reynolds number (~10^5) than larger transonic commercial aircrafts (~10^8), due to their reduced size and operating velocity. Inertial forces are still dominant in the flow at this Reynold number range, which is the reason why turbulence still needs to be considered in the computational problem. This modelling is achieved in CFD software through turbulence closure schemes. In fact, *Spalart-Allmaras* (SA) and *k-ω SST* are two turbulence models which were applied to external aerodynamics problems. For instance, Panagiotou et al. [3] used *SA* in their UAV aerodynamic analysis, while Kontogiannis et al. [4] employed *k-ω SST*. The versatility between the two models is due to the fact that the former was specifically developed for the aerospace industry [5], and the latter is known to have satisfactory performance in transitional and low Reynolds number flows [6]. As the UAVs from the above-mentioned studies featured a more traditional wing configuration, the present study employs both of these two turbulence models to evaluate their performance on a UAV with a canard wing configuration, whose respective results were then compared against validation data obtained with the semi-empirical XFLR5 software.

Studies by Kanesan et al. [7] and Ramos [8] have verified FEA predictions of the structural performance of winds to external forces. The present study adopted the same procedures outlined in these previous investigations due to the shared similarities including boundary conditions and material selection.

The FSI analysis for different flight conditions—23 and 13 m s^{-1} for *wind gust* and *normal* flight conditions, respectively—provides sufficient information to predict the structural performance of the UAV wings and wing-fuselage connectors, ensuring that a sudden wind gust would not lead to a catastrophic in-flight failure of the UAV.

2 Methodology

The following section presents the present study's methodology to perform a FSI simulation: first, flow simulations are obtained through CFD, which are then coupled with structural analysis through FEA. The methodology to predict the stresses in the wing-fuselage connectors through analytical calculations is also explained last.

2.1 Fluid Simulation Using CFD

The procedure that is employed to carry out the flow simulation is now described. A half-body computational domain is used due to symmetry in the geometry of the wings, as shown in Fig. 4 to enhance the mesh resolution while balancing computational cost. The generated domain is made of approximately 3 million cells (Fig. 5), with inflation layers having a $y^{+} = 1$, in order to guarantee that the boundary layer phenomenon is captured adequately as described in previous studies [3, 4, 9, 10].

Table 2 summarises the employed boundary conditions for the flow simulation. Two flight-conditions: *normal* and *wind gust* are studied corresponding to two different inlet velocities. The resultant pressure fields are used in the subsequent structural analysis. Ultimately, the goal is to determine whether the UAV wings will be able to withstand the additional stresses introduced by an increase in airspeed.

The flow field is determined with the *SA* and *k-ω SST* turbulence models, which was already presented earlier in this study. The aerodynamic wing loading is obtained through the numerically calculated pressure distribution.

To validate the CFD results, the same geometry and boundary conditions are tested using the semi-empirical CFD software XFLR5 in order to compare the numerical results against those obtained from ANSYS. XFLR5 is considered a viable alternative in the absence of experiments. The software has been benchmarked against experimental results in the past with overall satisfactory prediction [11, 12]. A separate study has not only demonstrated XFLR5's capability to accurately simulate low Reynolds numbers flows but has also shown its widespread use in flow simulations regarding a range of airfoil-dependent investigations [13].

2.2 Structural Analysis

Composite materials are used to manufacture the wings as shown in the CAD model in Fig. 6. The back-wing comprises of a EPP foam core shelled by a 3 mm thick

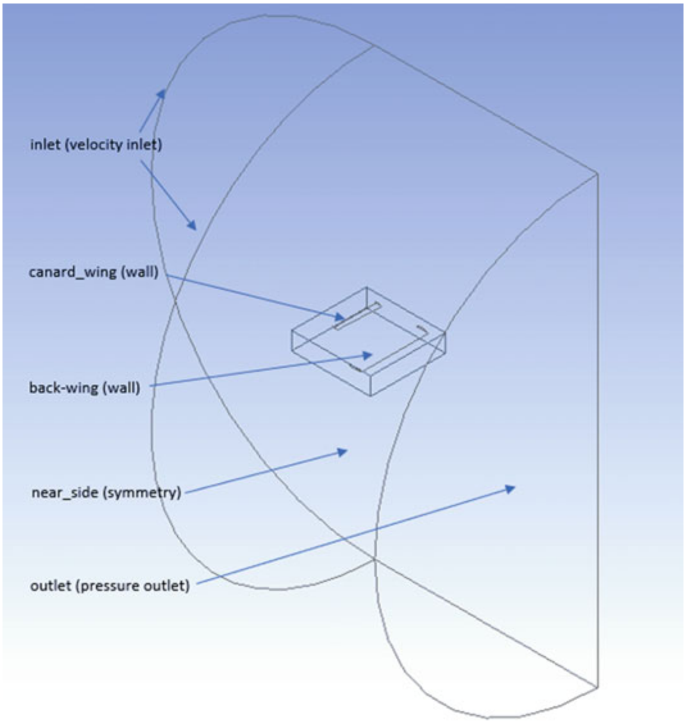


Fig. 4 Computational domain employed for flow simulation in ANSYS Fluent (CFD)

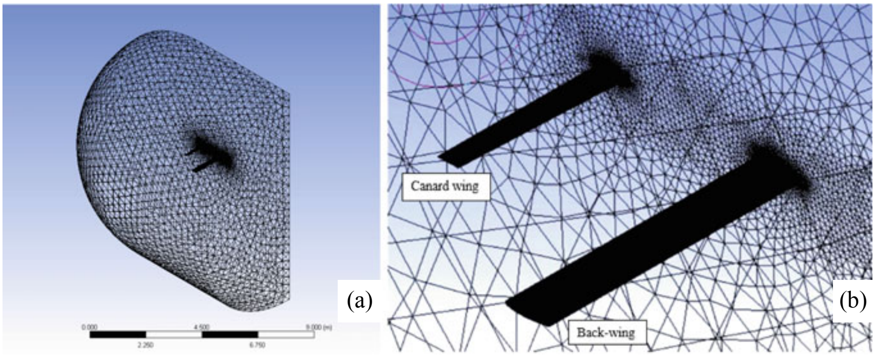


Fig. 5 CFD computational grid: **a** isometric view; **b** close-up view of mesh refinement in the vicinity of the wings

Table 2 Boundary conditions for flow simulations

Boundary name	Boundary type	Boundary conditions
Inlet	Velocity inlet	$p = 0$ atm (gauge pressure); $T = 300$ K; $v_1 = 13$ m s^{-1} (*normal state*); $v_2 = 23$ m s^{-1} (*wind gust*); Angle of incidence $= 3.0°$
Near side	Symmetry	Symmetrical with respect to boundary
Canard wing; Back-wing	Wall	$v = 0$ m s^{-1} (no-slip condition)
Outlet	Pressure outlet	–

Cf. Fig. 4 for the various boundary locations

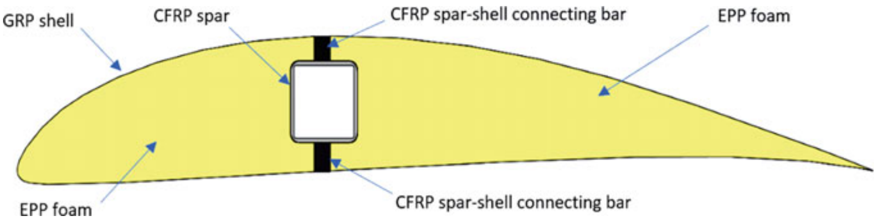

GRP shell
CFRP spar
CFRP spar-shell connecting bar
EPP foam
EPP foam
CFRP spar-shell connecting bar

Fig. 6 Back-wing's wingroot cross-section—by Haggis Aerospace [1] but owned by the author K. Marangi

hand-laminated GRP, with a CFRP tube running through the location where the wing is at its thickest. The CFRP tube acts as the wing's spar and is designed to absorb the aerodynamical loads transmitted from the outer part of the wings. Unidirectional CFRP bars are employed to convey the torsional moment from the wings' skin to the CFRP tubes [14].

Pressure fields are extracted from the flow simulations in ANSYS FLUENT and are then mapped onto the wings' equivalent structural models in ANSYS MECHANICAL. This stage is where the coupling between CFD and FEA occurs, or in other words, where the fluid flow problem is converted into a structural one.

The anisotropic mechanical properties of the various materials illustrated in Fig. 6 (Young's Modulus Ex Ey Ez & Poisson's ratio ν_{xy}, ν_{yz}, ν_{xz} are listed in Table 3. These are extracted from ANSYS' Material Library [15].

2.3 Analytical Analysis

The back-wing connectors (Fig. 2) are modelled with a number of assumptions to facilitate the analytical calculations of the resulting stresses. The connectors comprise of two components: the 3D printed parts and the M4 bolts (cf. Table 1). The following assumptions are made for the analytical calculations:

Table 3 Mechanical properties of wings' materials

Parameter	Glass fibre Reinforced Plastic (woven, wet)	Carbon Fibre Reinforced Plastic (woven, wet)	Carbon Fibre Reinforced Plastic (UD, wet)
Ex (MPa)	35,000	59,160	1.23×10^5
Ey (MPa)	9000	59,160	7780
Ez (MPa)	9000	7500	7780
ν_{xy}	0.28	0.04	0.27
ν_{yz}	0.4	0.3	0.42
ν_{xz}	0.28	0.3	0.27

Table 4 Mechanical properties of the aluminum alloy bolt, assuming material homogeneity and isotropy

Parameter	Value
Young's Modulus (GPa)	72
Yield strength (MPa)	505

- The front connectors are located above the wing's neutral point (the entire generated aerodynamic load is applied to this fixed region);
- The aerodynamic load is directly applied to the connectors nutted bolts, which provide the necessary clamping power;
- The drag force is ignored because it is negligible compared to the lift force experience;
- The mechanics of the fasteners are not considered, and the stress limits of the bolt material [16] is set as the determining factor (assuming homogeneity and isotropy of the bolt) (Table 4).

Based on these assumptions, it is possible to calculate the resultant tensile stresses in the connectors M4 bolts for the two different flight conditions using the Formula (1) [where σ is the resulting stress (Pa), F is the applied force (N), & A the item's cross-sectional area (m^2)].

$$\sigma = \frac{F}{A} \tag{1}$$

3 Results and Discussions

This section presents the results from the FSI simulations, starting with the CFD and FEA results, then ending with the analytical calculations.

3.1 Flow Simulation

The generated aerodynamic loads (lift and drag) from the CFD simulations are presented in Table 5.

For the *normal* wind condition, ANSYS predicted a sensible lift force as one would expect in scenarios where the UAV reaches level-flight. This implies that the lift and the weight - two vertical opposing forces—are cancel each other out. For the normal flight condition, the total lift forces are 77.2 N using *k-ω SST*, and 77.8 N using *SA*. It is possible to partially validate the CFD results from the ANSYS-predicted lift forces as these correspond to typical level-flight cases for this type of UAV.

For the *wing gust* condition, *k-ω SST* and *SA* obtained similar total lift values of 244 N and 246 N, respectively. These values imply that there has been an increase of 216% in aerodynamic forces due to the sudden acceleration experienced by the UAV. Nevertheless, this significant rise in lift is expected, as the vehicle's velocity is a squared term in the lift Formula (2) [where ρ is the air's density (Pa), v the airspeed (ms^{-1}), S the projected wing area (m^2), and C_L the lift coefficient].

$$L = \frac{1}{2}\rho v^2 S C_L \tag{2}$$

Regarding the different values presented in Table 5, both turbulence closure schemes obtained similar lift results, with less than 1% error in both flight conditions. In fact, there is a 0.78% error between *k-ω SST* and *SA* for *normal* conditions and 0.82% for *wind gust* conditions. These minor discrepancies imply that either turbulence model is suitable to their similar performance.

To validate the numerical results, the numerical results produced by *k-ω SST* and *SA* in ANSYS FLUENT are compared against values generated by XFLR5, which is displayed in Table 6 for the aerodynamic coefficients. From this table, the drag coefficient is unanimously predicted to have a value of 0.05, although there is an approximate 20% difference between ANSYS and XFLR5 results in terms of lift coefficient. Taking into account that the lift coefficient is calculated from the lift Formula (2) shown above, this discrepancy could be explained by the fact that the

Table 5 Generated aerodynamic loads for the two flight conditions

Parameter	Value	
Normal conditions ($V = 13\ ms^{-1}$)		
Turbulence model	*k-ω* SST	SA
Lift force (N)	77.2	77.8
Drag force (N)	4.34	4.04
Wind gust conditions ($V = 23\ ms^{-1}$)		
Turbulence model	*k-ω* SST	SA
Lift force (N)	244	246
Drag force (N)	13.2	13.2

Table 6 Aerodynamic coefficients from CFD simulations

Parameter	Value		
	k-ω SST	SA	XFLR5[a]
Lift coefficient	0.96	0.97	0.76
Drag coefficient	0.05	0.05	0.05

[a] XFLR5 values obtained from the student team (Haggis Aerospace) in charge of the UAV [1]. Data owned by the author K. Marangi

other variables of the formula had different values when calculated in XFLR5. These differences could be related to the air viscosity or density.

Figures 7 and 8 provide qualitative results of the flow simulations for both flight conditions based on the results from *k-ω SST* (note that both turbulence models performed similarly). In these figures, the pressure coefficients across the wings' surfaces are depicted through contours. These coefficients describe the relative pressures throughout a specific flow field, based on Eq. (3) [where C_p is the pressure coefficient, p the static pressure at a specific location (Pa), p_∞ the freestream static pressure (Pa), p_0 the freestream stagnation pressure (Pa), ρ_∞ the freestream fluid density (kg m^{-3}) & V_∞ the plane's velocity (m s^{-1})].

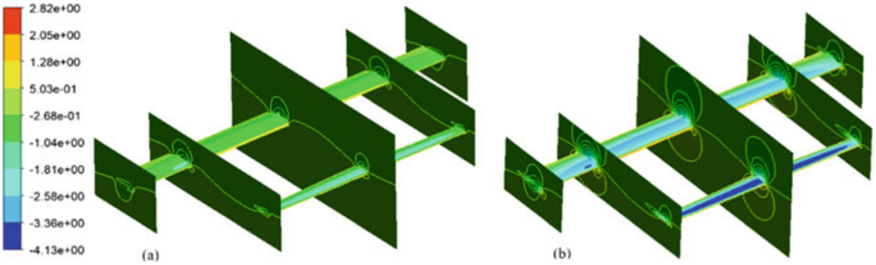


Fig. 7 Pressure coefficient contours for: **a** normal, 13 ms^{-1} and **b** wind gust, 23 ms^{-1} flight conditions

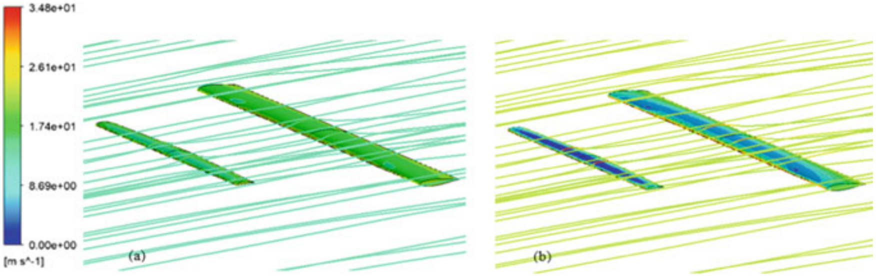


Fig. 8 Streamlines and pressure coefficient contours for: **a** normal 13 ms^{-1}, and **b** wind gust 23 ms^{-1}, flight conditions

$$C_p = \frac{p - p_\infty}{\frac{1}{2}\rho_\infty V_\infty^2} = \frac{p - p_\infty}{p_0 - p_\infty} \tag{3}$$

In Fig. 7, the contours generally show low values of pressure coefficient on the upper wing surfaces compared to higher values on the lower surfaces. An upward lifting force is generated as a result of this pressure difference. Comparing the *wind gust* (Fig. 7b) against the *normal* conditions (Fig. 7a), a larger pressure difference is observed and as a consequence greater aerodynamic loads are experienced. This agrees with the information found in Tables 5 and 6. Moreover, Fig. 8 demonstrates that the airflow on the upper wing surfaces has the largest velocity, which agrees with Bernoulli's principle: the faster the flow, the lower its pressure.

The flow simulation results from ANSYS CFD, specifically the pressure fields, are exported to ANSYS Mechanical to carry out the second step.

3.2 Structural Analysis

The structural investigation is executed in two stages. Firstly, static FEA simulations in ANSYS Mechanical predict the wing stresses due to aerodynamic loads. In the second step, the reaction forces obtained in the FEA simulation are used to analytically calculate the resultant stresses induced at the wing-fuselage connectors from the aerodynamic loads.

Table 7 summarises the results from ANSYS Mechanical for each of the wings during both flight conditions.

The largest magnitude of the maximum stress is experienced on the back-wing, with values of 75.0 MPa during *wind gust* and 23.8 MPa during *normal* condition.

Figure 9 illustrates the maximum equivalent stresses. The largest stresses are concentrated on the back-wing spar-shell connecting bar (labelled 'max' on Fig. 9b).

Table 7 Results from ANSYS mechanical

Parameter	Value	
Wing	Canard	Back
Flight conditions	*Normal* ($v_1 = 13$ ms^{-1})	
Maximum Stress (MPa)	3.30	23.8
Reaction Force (N)	13.5	25.0
Maximum Deformation ($\times$ 10^{-3} m)	3.00	0.70
Flight conditions	*Wind gust* ($v_2 = 23$ ms^{-1})	
Maximum stress (MPa)	10.0	75.0
Reaction force (N)	42.2	78.9
Maximum Deformation ($\times$10^{-3} m)	9.50	2.25

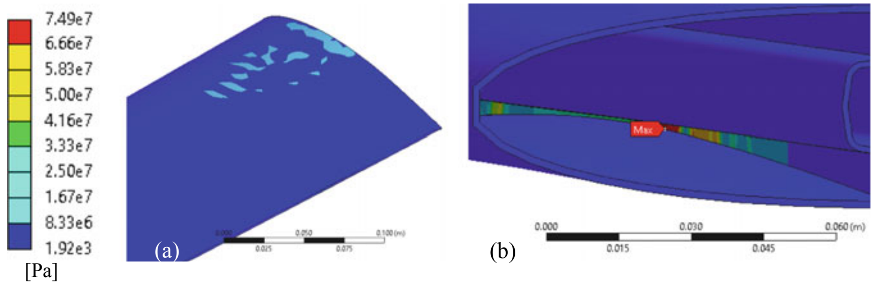


Fig. 9 Equivalent stress contours of **a** canard wing and **b** back-wing (close-up view on regions of high stresses for each wing)

To determine whether these high stresses pose a risk of damage the maximum stress failure criteria was employed. The maximum stress failure criteria is based on non-interactive theory [17], which does not take consider the interaction between the different composite elements. The computed stress is then evaluated against the component's material stress limit. The computed stresses are compared to the orthotropic stress limits and are below the tensile limit of 1632 MPa and compressive limit of −704 MPa (cf. Table 3).

The ANSYS Mechanical reaction forces for each flight condition are exported as inputs in the tensile stress Formula (1) for the investigation of the transmitted stresses to the wing-fuselage connector. These are presented in Table 8. The M4 bolt experiences a tensile stress of 1.98 MPa during *normal* conditions, increasing to 5.95 MPa during the *wind gust*. These stresses are below the bolt yield strength of 505 MPa, producing a factor of safety of 85. This implies that the implemented connector design is satisfactory and will not fail under the assumed flight conditions.

Finally, Fig. 10 illustrates the resultant deformation for the two flight conditions. From Table 7, it is observed that the canard wing experiences the largest deflection with a value of 3.00 mm during *normal* flight condition, increasing to 9.5 mm during the *wind gust* (an increase by a factor of 3). This could be explained by the fact that the canard lacks a component connecting its GRP shell to its spar, and therefore the shell takes on the entire stresses from the aerodynamic lifting force.

Table 8 Results from Analytical analysis

Parameter	Value	
Flight condition	Normal ($v_1 = 13\ \text{ms}^{-1}$)	Wind gust ($v_2 = 23\ \text{ms}^{-1}$)
Cross-sectional area (m^2)	1.26×10^{-3}	
Applied Force (N)	25.0	78.9
Resultant Stress (MPa)	1.98	5.95

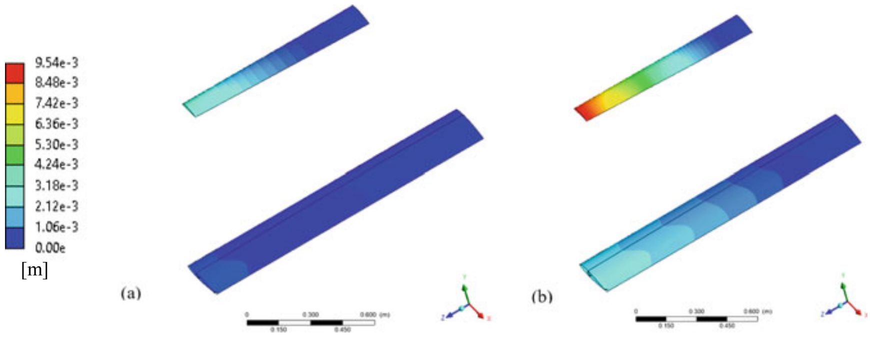


Fig. 10 Total deformation contours on wings for: **a** normal (13 ms^{-1}) and **b** wind gust (23 ms^{-1}) flight conditions

4 Conclusion

The present study examines the structural integrity of the wings and wing-fuselage connectors of a UAV exposed to varying aerodynamic loads during two different flight conditions using Fluid-Structure Interactions. ANSYS FLUENT is first used to numerically simulate the airflow around the wings. The resultant pressure fields are imported into ANSYS MECHANICAL to carry out a static structural analysis (FEA). The wings' deformation and maximum stresses due the aerodynamic loads are predicted. Lastly, the stresses in the back-wing's wing-fuselage connector are calculated analytically, and the maximum stress failure criteria is applied.

It is observed that the pressure differential between the upper and lower surfaces of the wings is much larger during *wind gust* compared to *normal* condition due to the increase in airspeed. Consequently, the aerodynamic forces increased as well, leading to significant larger stresses and to deformations (by a factor of 3). However, these resultant stresses were still below the limits by a very safe margin.

The FSI simulations that were presented in this study can be applied to related studies as it allows engineers to evaluate the aerodynamic and structural performance of a mechanical system without the need of a physical prototype, which is often costly and time-consuming to implement.

Acknowledgements We thank the University of Dundee's engineering student team participating in the *IMechE UAS Challenge* competition—HAGGIS AEROSPACE—and Dr. Triantafyllos Gkikopoulos (RaptorUAS) for suggesting this research title and for the insightful discussions on UAVs. We would like to also thank the Institution of Mechanical Engineers for supporting us in the participation of the IAENG 2019 conference.

References

1. Haggis Aerospace, *Critical Design Review*. Document submitted for IMechE UAS Competition 2017, Dundee (2017)
2. K. Marangi, S. M. Salim, Predicting the structural performance of the wings of an unmanned aircraft vehicle using fluid-structure interaction, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, London, UK, 3–5 July 2019, pp. 500–505
3. P. Panagiotou, P. Kaparos, C. Salpingidou, K. Yakinthos, Aerodynamic design of a MALE UAV. Aerosp. Sci. Technol. **50**, 127–138 (2016)
4. S. Kontogiannis, D. Mazarakos, V. Kostopoulos, ATLAS IV wing aerodynamic design: from conceptual approach to detailed optimization. Aerosp. Sci. Technol. **56**, 135–147 (2016)
5. P. Spalart, S. Allmaras, A one-equation turbulence model for aerodynamic flows, in *30th Aerospace Sciences Meeting and Exhibit* (1992)
6. S. Kontogiannis, J. Ekaterinaris, Design, performance evaluation and optimization of a UAV. Aerosp. Sci. Technol. **29**(1), 339–350 (2013)
7. G. Kanesan, S. Mansor, A. Abdul-Latif, Validation of UAV wing structural model for finite element analysis. Jurnal Teknologi **71**(2) (2014)
8. M. Ramos, Construction and analysis of a lightweight UAV wing prototype, M.S thesis, Técnico Lisboa, Lisbon, Portugal (2015)
9. S.M. Salim, S.C. Cheah, Wall y^+ strategy for dealing with wall-bounded turbulent flows, in *Proceedings of International MultiConference of Engineers and Computer Scientists*, vol. 2 (2009), pp. 2165–2170
10. S.M. Salim, M. Ariff, S.C. Cheah, Wall y^+ approach for dealing with turbulent flows over a wall mounted cube. Prog. Comput. Fluid Dyn. **10**(5–6), 341–351 (2010)
11. A. Deperrois, About XFLR5 calculations and experimental measurements (2009). [ebook]. Available http://www.xflr5.com/docs/Results_vs_Prediction.pdf
12. J. Morgado, R. Vizinho, M. Silvestre, J. Páscoa, XFOIL vs CFD performance predictions for high lift low Reynolds number airfoils. Aerosp. Sci. Technol. **52**, 207–214 (2016)
13. XFLR5, Xflr5.com (2018). (Online). Available http://www.xflr5.com/xflr5.htm. Accessed 2 Nov 2017
14. M. Simons, *Aerodynamics of model aircraft flight*, 5th edn. (Special Interest Model Books Ltd, Dorset, England, 2015)
15. Workench Mechanical. Ansys
16. Aluminum Socket Cap Bolt M4 x (0.7 mm) x 20 mm, Pro-bolt.com, 2015. (Online). Available https://www.pro-bolt.com/aluminium-allen-bolt-m4-x-0-7mm-x-20mm-21.html. Accessed 5 Mar 2018
17. N. Tiwari, *Introduction to Composite Materials and Structures: Strength of a Composite Lamina* (2018). [ebook] Available https://nptel.ac.in/courses/112104168/L32.pdf

A Vision Guided Robot for Gluing Operations

Stefano Pagano, Riccardo Russo, and Sergio Savino

Abstract The paper describes the development of an automatic machine to be adopted to glue the shoe upper to its sole or to glue a rubber insert on the lower surface of the sole. The machine prototype consists in a cartesian robot that drives a glue gun and by a vision system that can recognize the sole, placed on a worktop, allowing the planning of the glue gun trajectory. After a description of the machine hardware assembly, the developed procedures that allows the sole recognition and the robot planning trajectories are presented. Finally, the results of several tests, performed to check the procedure goodness, are reported.

Keywords Arduino · Automatic machine · Footwear manufacturing · Gluing process · Soles · Vision sensor

1 Introduction

In the last 50 years, shoes have undergone considerable evolution, both in terms of materials and fixing systems for their parts; in a recent past, the soles were mainly made of leather, sometimes with rubber inserts placed on the lower surface and the upper was fixed to the sole by means of hooks and seam.

Currently, the large series production provides shoes made with increasingly lighter materials and equipped with soles made of deformable material that are, at the same time, comfortable, breathable, durable and able to provide good thermal protection. Great importance is given to the sole cushioning function, especially for sports shoes that, in some cases, are equipped with gas-filled bag, inserted in the sole

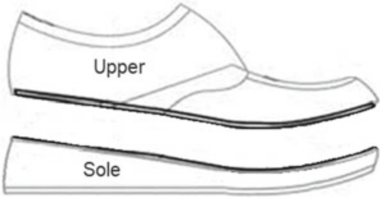
S. Pagano (✉) · R. Russo · S. Savino
Dipartimento di Ingegneria Industriale, Università degli Studi di Napoli Federico II, via Claudio n.21, 80125 Naples, Italy
e-mail: pagano@unina.it

R. Russo
e-mail: riccardo.russo@unina.it

S. Savino
e-mail: sergio.savino@unina.it

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_2

Fig. 1 Shoe upper and sole



to prevents impacts [1]. Therefore, in addition to traditional leather soles, there is a vast production of synthetic materials soles belonging to the following groups or their combination [2, 3]:

- thermoplastic rubber (TPR) and thermoplastic polyurethane (TPU);
- two-component polyurethane materials: polyether-based PUR, polyester-based PUR;
- copolymers such as rubber and EVA (Ethylene vinyl acetate).

The soles for footwear must comply with the requirements indicated by various international standards to be labeled as quality sole.

The standard establishes the tests to be performed to assess resistance to bending, abrasion, de-lamination, slip, water penetration, dimensional stability, compressive and splitting tensile strength, of the stitching point and to bonding capacity.

When uppers and soles (Fig. 1) are joined by gluing, the junction zone is subject to a combination of tensile, shear and peel stresses; the most critical stress is the peel one. For this reason, special instruments are adopted to test the adhesion strength between upper and sole. The load causing the separation can be measured or alternatively, a pass load can be applied to check that the adhesion is satisfactory. This second operation method is the more adopted in the shoe factory since it can be applied to the ordinary shoes of the production lines.

In the standard EN 15307 there are reported the minimum shoes peel strength for different kind of shoe. For example, it must be greater than 3.0 N/mm^2 for men town footwear; 2.5 N/mm^2 for women town footwear; 5.0 N/mm^2 for mountain footwear.

This paper refers about the development of a gluing machine prototype, called Ulisse [4] that allows to connect uppers and rubber inserts to the soles; the gluing operation must be carried out with great care in order to meet the current standards of resistance; at the same time, it must be carried out as quickly as possible to contain production costs. The automatic machine prototype uses a vision system to recognize the sole on which the glue must be dispensed, allowing to plan the trajectory of a cartesian robot that drives the glue gun. It was implemented with procedures that allow to compensate for mechanical misalignments between robot and vision system and to identify some particularities of the soles such as tapered edges or raised edges [5]. Some preliminary tests are reported in [6].

The developed prototype improves all the parameters of the production tetrahedron. In fact, it allows to improve:

(a) the production *costs*. The machine may be produced by means of cheap preassembled components and does not require operators having specific skills;
(b) production *time*. The developed procedures allow to arrange the soles to be glued in an arbitrary position on the worktop and does not require preliminary operations such as storing the sole geometry;
(c) production *flexibility*. The prototype can operate on soles having different sizes and shapes;
(d) production *quality*. It is possible to control the amount of glue dispensed by adjusting the glue gun velocity, the distance from the object and the glue flow rate, allowing to guarantee the required adhesion resistance.

2 Prototype Description

The gluing machine prototype (Fig. 2) was made with pre-assembled and low-cost elements. It mainly consists of a closed cabin containing a worktop on which are placed the soles to be glued; a vision system able to recognize the shape of the soles; a Cartesian robot whose end-effector is a glue gun; an aspirator, placed under the worktop, to suck the glue vapors avoiding their dispersion in the work environment. The main components of the machine are described in detail below.

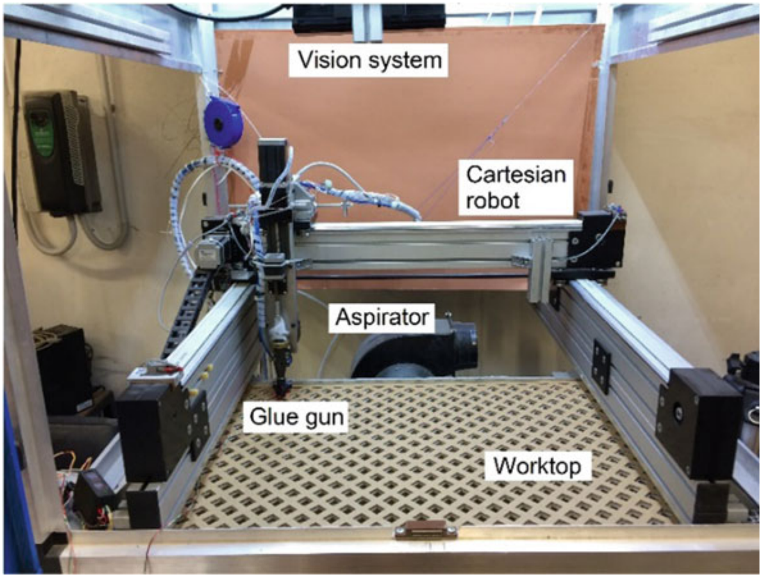


Fig. 2 Gluing machine prototype

(A) *The cabin*

The cabin frame was realized by means of a modular system of standard aluminum components which allow a simple and fast assembly. It also allows to modify the structure geometry to adapt it to the variations that arise during the prototype set-up activity and to adjust its stiffness so that it can bear the robot weight preventing vibrations induced by the rapid robot movements.

(B) *The worktop*

The worktop, on which are placed the soles to be glued, has an area of 550 × 650 mm; it is constituted by a permeable plane so that the suction action, carried out by an aspirator located under the worktop, may convey and remove the glue vapors from the cabin. For this reason, the worktop consists of a pressed cardboard grid, covered with a fabric (not present in Fig. 2); both elements must be periodically replaced as they can be impregnated by the glue vapors crossing them.

The suction action also realizes a restrain function for the soles placed on the worktop.

(C) *The robot*

The cartesian robot is placed in the cabin; it has the task of driving the glue-gun. To reduce the production costs [7], the cartesian robot was made of pre-assembled elements easily available on the market and ready-to-install. The translational motions, along the three orthogonal axes, are obtained by means of linear modules whose slides are composed of a steel plate translating on idler wheels. The horizontal slides are driven by a toothed belt transmission (Figs. 3 and 4), while the vertical slide, supporting the glue gun, is driven by a trapezoidal screw-nut transmission.

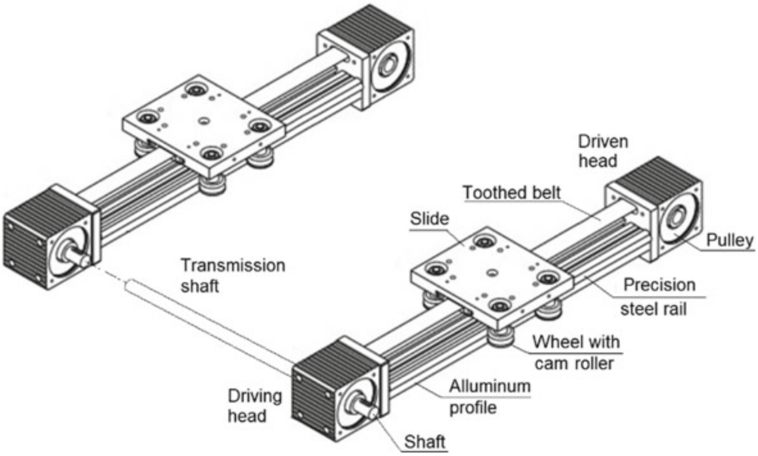


Fig. 3 Linear modules with toothed belt drive for the first translational motion

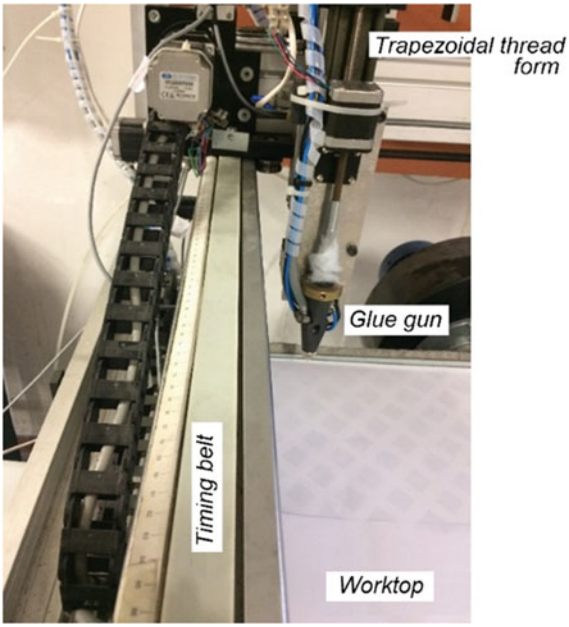


Fig. 4 Linear motion systems of the prototype

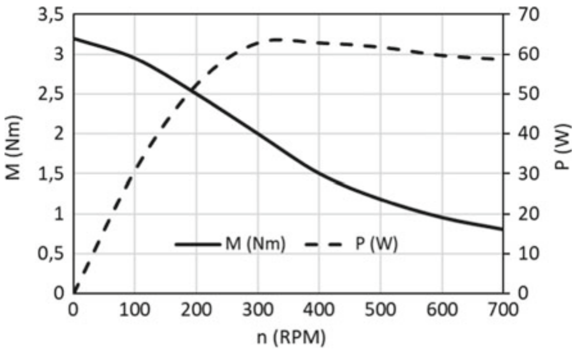
The belt transmissions allow to perform rapid planar movement with maximum speed and acceleration equal to about 3 and 7 m/s^2, respectively. The vertical positioning does not require rapid movements and high precision; for this reason, a screw-nut transmission is adopted; thanks to its low efficiency (30% or less), this transmission does not allow the slide to move downwards under the effect of its weight so that the end-effector can be maintained at a constant height without energy consumption. In any case, adopting a screw pitch of 2 mm and driving it with a stepper motor with a step-angle of 1.8°, the vertical displacement accuracy may be lower than 1/100 mm.

The assembly of the linear modules is simple and inexpensive but may involve errors in the end-effector positioning due to inaccurate components assembly, backlashes between the moving parts and deformability of some components.

To limit the positioning errors, some mechanical precautions must be provided:

- by means of the wheels eccentric pins it is possible to adjust the preload between wheel and rail in order to remove the backlash and to prevent the wheel-rail sliding; this operation minimizes the rolling friction thus improving the mechanical efficiency;
- to reduce the belts deformability, the driven pulleys of the planar linear modules are mounted on an eccentric pin allowing the belt tensioning; in this way it is possible to guarantee a positioning repeatability of ±0.1 mm;
- as the first translation motion is obtained by means of two parallel guides, the two driving pulleys are connected by means of a transmission shaft (Fig. 4) so that the same motor drives both the slides assuring the synchronization of the movements.

Fig. 5 Stepper motor characteristic curves (LAM Tech., mod. NEMA17: 3.2 Nm bipolar static torque, 4.2 Arms bipolar phase current, 24 Vdc)



The stepper motor adopted for the three translations can provide a maximum torque of 3.2 Nm; torque and power curves are shown in Fig. 5. The glue gun nozzle is adjusted with a stepper motor whose maximum torque is equal to 0.5 Nm.

The robot is controlled by the microcontroller Arduino UNO board, managed by GRBL Software [8, 9]; it controls the motor drivers, the end-stroke of the linear modules and the glue gun opening/closing valves.

(D) *The vision system*

The 3D vision system allows the recognition of objects to be glued, in order to detect their position on the worktop, their dimensions and the contour along which the glue must be dispensed. The Ulisse prototype is equipped with a Microsoft Kinect V2 that is a camera, integrated with an infrared sensor, that allows to measure the sizes of the framed objects [10]; the vision system is positioned on the ceiling of the cabin to frame the entire worktop. It consists of an RGB camera, whose resolution is 1920 × 1080 pixels, and an infrared emitter combined with an IR camera with a resolution of 512 × 424 pixels; the elaboration of the IR signals allows to define the depth map.

The vision system reliability was checked for objects placed at distances from the camera included in the range 0.50–0.80 m; the tests stated that it is reliable and repetitive from 0.60 m; it was therefore fixed on the cabin ceiling at the distance of 0.75 m from the worktop.

3 Calibration Procedure

Two fixed frame references are defined for the prototype Ulisse: the first one (*Rr*) has its origin in the robot base; the other one (*Rc*) in the vision system camera. The two frames have different origin position and different orientations. To transform the coordinates detected by the vision systems from *Rc* to *Rr* reference, it must be defined a homogeneous transformation matrix [T]:

$$\{x\}_r = [T] \cdot \{x\}_c \tag{1}$$

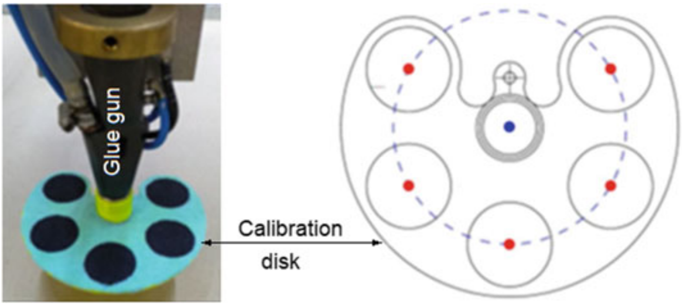


Fig. 6 Disk adopted for the calibration procedure

being: $\{x\}_r$ and $\{x\}_c$ the coordinate vectors expressed in the *Rr* and *Rc* references, respectively. The homogeneous matrix:

$$[T] = \begin{bmatrix} C_{11} & C_{12} & C_{13} & V_x \\ C_{21} & C_{22} & C_{23} & V_y \\ C_{31} & C_{32} & C_{33} & V_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2}$$

contains the C_{ij} elements of the rotation matrix and Vi elements to define the *Rc* origin position in *Rr* reference. The matrix contains six unknowns: three rotations and three displacements.

The calibration procedure allows to identify this matrix by means of the following procedure that adopts a tool that allows the system vision to identify the position of the glue-gun nozzle. This tool is constituted by a plastic disk that can be fixed on the glue gun with its centre coincident with nozzle centre (Fig. 6). The disk is covered with a cyan fabric on which five circles, made of black fabric, are placed. The centres of the black circles are arranged on a circumference whose centre coincides with the centre of the cyan disk and therefore with the position of the glue gun nozzle centre.

The calibration procedure requires the positioning of the glue gun nozzle in several different points of the volume framed by the camera. In each position the camera recognizes the visible black circles (at least three); by adopting classical computer vision algorithms [11–14], the positions of the centres of the black circles are first identified and subsequently the position of the centre of the circumference passing through these points is defined that, as said above, coincides with the glue gun nozzle position.

For each robot configuration, the nozzle position is detected in both, *Rr* and *Rc* references; by means of an optimization algorithm the transformation matrix is identified.

4 Robot Planning Trajectory

Starting from the recognition of objects to be glued, a data processing procedure has been developed in order to implement the trajectory that the glue gun must track.

The gluing of an upper shoe to its sole consists in depositing the glue along the contour of the sole upper surface that often differs from the sole top view contour (image contour) since the soles can have a tapered shape (Fig. 7).

To define the gluing contour, the three-dimensional information provided by the points cloud matrix, are used.

Starting from points cloud matrix D, containing the coordinates of the point in the *Rc* reference, the procedure transforms the coordinates in the *Rr* reference, obtaining the D1 matrix.

Then, to consider only the cloud points related to the object, the points belonging to a predefined volume, are selected. The volume is defined by a base surface equal to the worktop and height of 150 mm; it is slightly raised form the worktop so that even the points belonging to the worktop are discharged. With this operation, the D1 matrix is converted in a new logical matrix S that, for each point, contains "true" value if the point belongs to the parallelepiped, "false" value in the contrary case. As the camera frames the worktop from the cabin ceiling, this last matrix can be seen as a B/W image that allows to identify the "top view" of the object.

By means of the classical image processing algorithms, from the S matrix it is possible to extract the object silhouette contour. Moreover, it is possible to evaluate the local slope of the points in the neighbor of the object contour, by computing the local normal vectors [12].

The system must discharge the points having high local slope because they belong to the lateral tapered surface of the sole where the glue must not be laid (Fig. 7). Therefore, only the points arranged on the upper surface of the sole are considered to identify the upper surface contour; to drive the glue gun, the contour coordinates are moved toward the inside of the surface and fitted with a continuous function, in order to obtain a gluing trajectory that is slightly inside the defined contour. The glue gun trajectory is, therefore, obtained by means of an "offset" of the contour line to prevent the glue gun sprays on the lateral tapered surface.

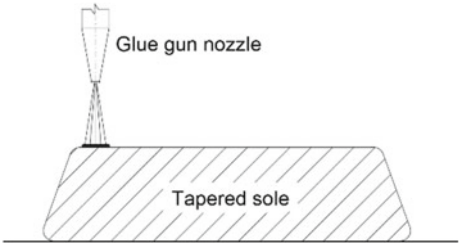


Fig. 7 Cross-section of a tapered sole

The processing time for the trajectory calculation dependents on the hardware performance and is independent on the number of objects to be analyzed. The prototype hasn't real-time procedures, but the developed algorithms are executed with waiting times that are compatible with production.

5 Experimental Results

The procedure was tested adopting a sample having a frustum of a cone shape (Fig. 8).

The sample was positioned on the worktop and framed by the camera (Fig. 9). The detected cloud points (Fig. 10) were elaborated to define the contour of the upper surface of the sample (blue line in Fig. 9) which is fitted with a continuous function (red line in Fig. 10).

To highlight the machine behavior, of the nozzle, a led laser was placed in correspondence of the glue gun nozzle, whose laser beam simulates the glue spray. The gluing machine is observed during the execution of a gluing test by means of another

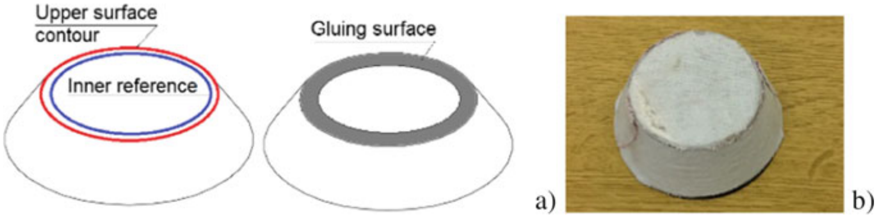


Fig. 8 Sample adopted to test the procedure: **a** reference entities, **b** the sample (D = 70 mm; d = 50 mm; h = 30 mm)

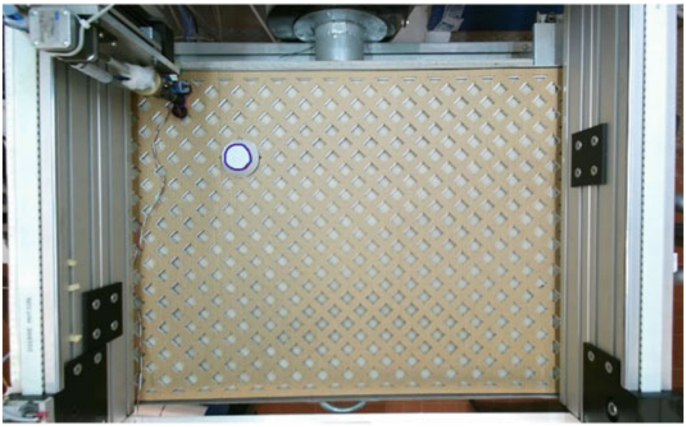


Fig. 9 Elaborated camera image containing the upper surface contour of the sample (blue line)

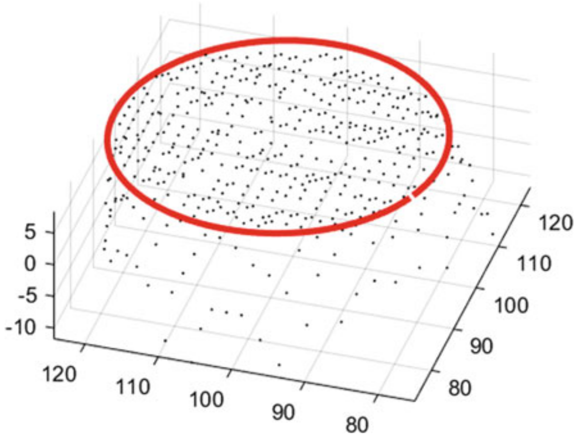


Fig. 10 The cloud points and the estimate contour of the upper sample surface

fixed vision system to record the video of the luminous point path. The laser beam path was isolated performing an analysis on successive frames, based on the normalized cross-correlation of matrices [15]. In particular, by means of correlation coefficient, a target sub-matrix of a frame "i", containing the laser mark on the object, was searched in the successive frame "i + 1", the submatrix found becomes the target for the following frame "i + 2"; this procedure was repeated for the whole video sequence; the video elaboration allows to visualize the glue path, as shown in Fig. 11 where the laser beam produces a mark on the object characterized by a strong brightness central area and a red halo.

The procedure was then tested on a shoe sole, as represented in Fig. 12. In this case the gluing surface is comprised between the raised sole contour and the lightened area of the sole.

The adopted procedure allows to identify the outline of the sole; then, through the evaluation of the local slope of the outline points, it excludes the outermost part (which has a raised border) and the lightened internal one. The trajectory of the glue gun was defined by offsetting the contour towards the inside of the sole.

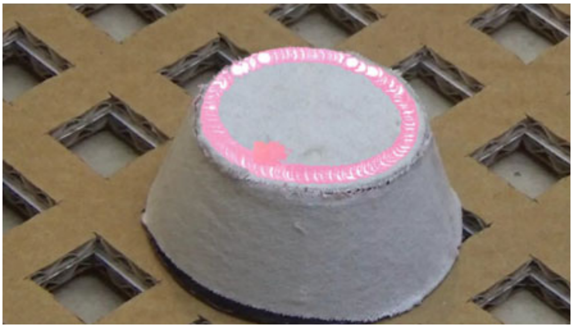


Fig. 11 Video elaboration allowing the visualization of the glue path

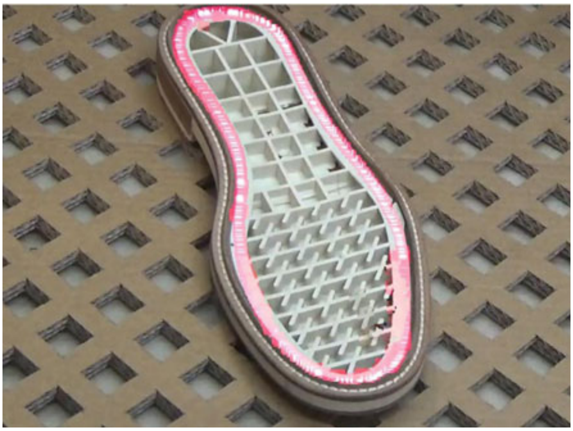


Fig. 12 Glue gun trajectory highlighted by means of a laser light connected to the glue gun nozzle

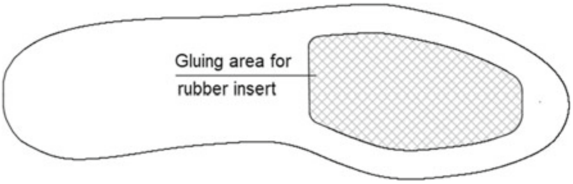


Fig. 13 Area for rubber insert in the lower surface of a leather sole

Other experimental tests were performed to evaluate a feature of the 2D procedure that allows to glue an anti-slip rubber insert in a depressed zone of the lower surface of a leather sole (Fig. 13). In this case the procedure must detect a contour placed inside the outline of the sole to plan a gluing trajectory with the aim of dispensing the glue in entire area delimited by the detected contour. These tests were conducted adopting the more performing vision sensor, *Intel Realsense D415* camera, characterized by a resolution 1280 × 720 pixels and by an RGB sensor with a resolution 1280 × 1080 pixels.

Therefore, the first part of the procedure allows to detect the sole outline; then, it searches an inside contour adopting the same image processing and edge extraction algorithms used to detect the external contour. The positive result of the process depends on the extension of the identified areas. To avoid the identification of small contours, like writings or drawings (Fig. 14), the identified contour must enclose an area whose extension must be greater than 1/5 of the area of the entire sole.

The procedure, tested on the leather flat sole (Fig. 14), was able to extract both the external and internal contour of the sole. Figure 15 shows the results of the image elaboration of the sole with the identification of the contours.

Finally, Fig. 16 reports the point cloud of the worktop with the sole placed on it; the same figure shows the internal contour (red) and the tridimensional glue gun trajectory (green), placed about 40 mm above the sole.

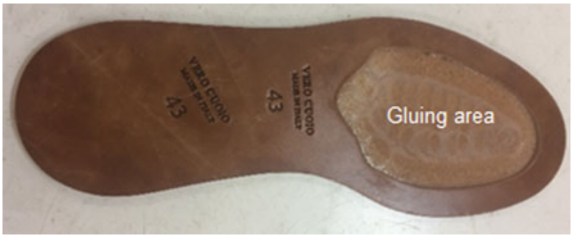


Fig. 14 Leather tested sole

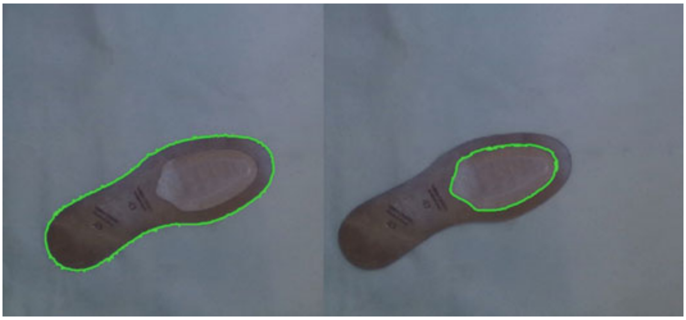


Fig. 15 Image elaboration: external and internal contour identification

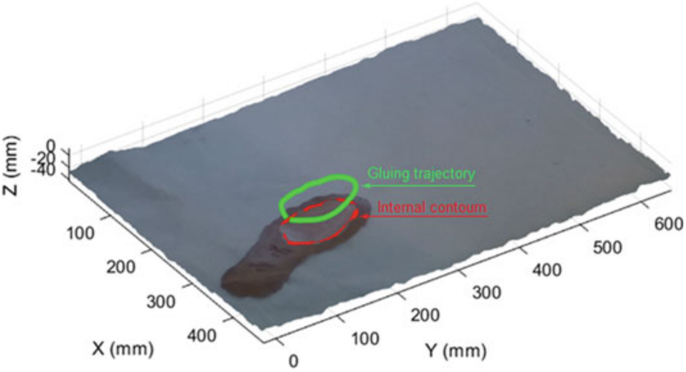


Fig. 16 Rubber insert contour and gluing trajectory

6 Conclusion

The development of gluing machine prototype has been presented; it has the possibility to recognize the sole to be glued by means of a vision system and to locate the gluing surface that is arranged in the neighborhood of the sole outline as it happens for the upper gluing operation or in an internal area for rubber insert gluing. This information is adopted to estimate the glue gun trajectory and therefore to drive the cartesian robot that move it.

The process is based on a preliminary calibration procedure that allows to identify the transformation matrix adopted to transform the coordinates detected by the vision systems in the robot reference.

The proposed procedure was verified by fixing a LED laser pointer to the glue gun nozzle and recording the track described by the luminous point projected on the gluing surface. This result confirms the goodness of the procedure that appear to be suitable for the footwear industry.

The adoption of a more performing hardware can produce an improvement in the performance of the machine regarding the execution time and the trajectory precision.

Acknowledgements The authors thanks Battista Accessori s.r.l., Michele Beato, Eliseo Malachia and Andrea Longobardi for their valuable support provided during this research activity; furthermore, they thank Giuseppe Iovino and Gennaro Stingo for their collaboration during the laboratory tests.

References

1. M.R. Shorten, D.S. Winslow, Spectral analysis of impact shock during running. Int. J. sport biomechanics **8**, 288–304 (1992)
2. The materials that create success soles for shoes. Available https://www.malaspina.it/en/materials-soles-for-footwear/
3. R.M.M. Paiva, E.A.S. Marques, L.F.M. da Silva, C.A.C. Antonio, F. Aran-Ais, Adhesives in the footwear industry. Proc. IMechE Part L J. Mater. Design Appl. IMechE 1–18 (2015)
4. L. Caruso, R. Russo, S. Savino, Microsoft Kinect V2 vision system in a manufacturing application. Robot. Comput. Integr. Manuf. **48**, 174–181 (2017)
5. S. Pagano, R. Russo, S. Savino, A vision guided robotic system for flexible gluing process in the footwear industry. Robot. Comput. Integr. Manuf. **65**. https://doi.org/10.1016/j.rcim.2020.101965, 2020
6. S. Pagano, R. Russo, S. Savino, Development and testing of a semiautomatic gluing machine. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019, 3–5 July, London, U.K. (2019), pp. 458–462
7. S. Pagano, C. Rossi, F. Timpone, A Low Cost 5 Axes Revolute Industrial Robot Design—11th Workshop on Robotics, RAAD 2002, Balatonfüred, Hungary, 2002
8. A. D'Ausilio, Arduino: a low-cost multipurpose lab equipment. Behav. Res. Methods **44**, 305 (2012)
9. S.S. Sarguroh, A.B. Rane, Using GRBL-Arduino-based controller to run a two-axis computerized numerical control machine, in *2018 International Conference on Smart City and Emerging Technology (ICSCET)* (2018)
10. A. Kolb, E. Barth, R. Koch, R. Larsen, Time-of-flight sensors in computer graphics, in *Proceedings of Euro Graphics* (2009), pp. 119–134
11. T.Y. Kong, A. Rosenfeld, *Topological Algorithms for Digital Image Processing* (Elsevier Science, Inc., 1996)
12. W.K. Pratt, *Digital Image Processing* (Wiley, 1991)
13. R.C. Gonzalez, R.E. Woods, S.L. Eddins, *Digital Image Processing Using MATLAB* (Pearson Prentice Hall, New Jersey, 2004)

14. S.M. Thomas, Y.T. Chan, A simple approach for the estimation of circular arc center and its radius. Comput. Vis. Graph. Image Process. **45**, 362–370 (1989)
15. J.P. Lewis, *Fast Normalized Cross-Correlation*, vol. 10 (Industrial Light & Magic, 2001)

Mechanical Modelling of Cylindrical Rings Versus Hollow Spheres Under Impact Loadings

Bridget Kogo, Matthew Farr, Bin Wang, and Mahmoud Chizari

Abstract This paper study caries a finite element simulation on metal rings subjected to fixed-end constant loading impacts against hollow spheres with close geometries. The mechanical parameters such as reaction force and strain were analysed and visual data recorded. Different rate of loadings and thicknesses were studied to examine the influence of these factors. The deformation processes of both rings and spheres were analysed and discussed, particularly the mechanism of deformation and load and strain histories for both rings and spheres were compared.

Keywords Finite element modeling · Impact loading · Mechanical testing · Ring · Spheres · Strain history

1 Introduction

Current study analyses a metal rings under end-impact (rather than axial-impact) crush loading, particularly looking at deformation characteristics. Due to metal rings' and tubes' high energy absorption capabilities as they undergo plastic deformation, they have a practical application in crash protection structures. Furthermore, as evidenced by Shim et al. [1], rings can be used as a simplified model for cellular material, allowing a range of biological and medical applications.

To The main aim and intention of the research is to gain an understanding of the deformation process and mechanism that a single ring undergoes when subjected to end loading. It is also intended to look at hollow spheres with similar cross-sections and modelling conditions and compare their deformation to that of a ring. The research will use Finite Element Analysis (FEA) software to examine this hypothesis.

B. Kogo (✉) · M. Farr · B. Wang
Department of Mechanical and Aerospace Engineering, College of Engineering, Design and Physical Sciences, Brunel University London, Uxbridge, UK
e-mail: biddyagada@yahoo.com

M. Chizari
School of Engineering and Computer Sciences, University of Hertfordshire, Hatfield, UK

School of Mechanical Engineering, Sharif University of Technology, Tehran, Iran

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_3

To simulate, a simple 3-body model will be used, whereby a large mass (modelled as a rigid wire) will be given an initial velocity to act as the impact load, and this will crush a metal ring, in turn supported by another wire at the distal end. These results will then be analysed and the model altered accordingly to test different variables, these being different impact velocities, as well as different radius-thickness ratios for the ring.

The actual deformation of a single metal ring under end impact has not been properly examined, although a mixture of experimental and theoretical analysis of deformation and stress wave propagation under impact loading exists for metal bars, rods and cylinders subjected to axial impact [2–4] and ring systems subjected to end impact [5–7]. Among the more significant studies are the latter, whereby a row of rings- in the same direction as the impact velocity- is either placed end-to-end or spot welded in place, in some cases with plates in between each ring to see how that effects the deformation of the system. The initial study in this field is Reid and Yella Reddy's [5]. Although an examination of the effect of inertia is the main aim of this study and [7], stress waves and their propagation formed a distinct part of both studies, as do the features of the modes of deformation.

Beyond studying the collapse of more complex cylinder systems, the ring systems studied in the article were simple, each being a single row of rings placed end to end on a horizontal plate, with the impact coming from a sledge.

Reid and Bell [8] improved on a previous model (Reid et al. [6], developed from [5]) by incorporating elastic waves, among other modifications. It's upon this basis that Reddy, Reid, and Barr's [7] attempted to build. The article introduced and adapted the previously developed methodology to be applied to free-ended systems in a manner similar to [5], also further refining the model developed through [6] and [8].

From [6], the use of high-speed cameras to record the deformation in visual detail was optimized so that the film could be analysed to more accurately calculate the varying velocity of the sledge. The rate of these cameras was also improved, producing 5000 frames per second. Further tests using strain-gauges on select rings on a different apparatus were performed, helping determine the speed that elastic waves propagate through ring systems.

Figure 1 shows examples of deformation in some of the ring systems. Unlike [5], this study focused on mild steel as the primary material, using brass for comparison in only 5/29 total tests ([5] used brass in 27/31 tests, in both annealed and 'as received' condition, with annealed mild steel for the remaining tests). The article also considered deformation in a single ring, attempting to model it from the pre-established theory.

One article examining the static effects on single rings being impacted is Reddy et al. [9], whose background relied heavily on the research published by Merchant [10], which defined and developed the concept of Equivalent Structures (Equivalent Structure Technique, EST, in [9]) to analyses plastic collapse problems.

Authors at [9] looked at rigid-perfectly-plastic materials and ones that underwent strain-hardening, a result of the initial plastic deformation. By considering the likely effects of strain-hardening, Reddy et al. built a more realistic model, particularly at the plastic hinges. This was evidenced by results found agreeing reasonably well

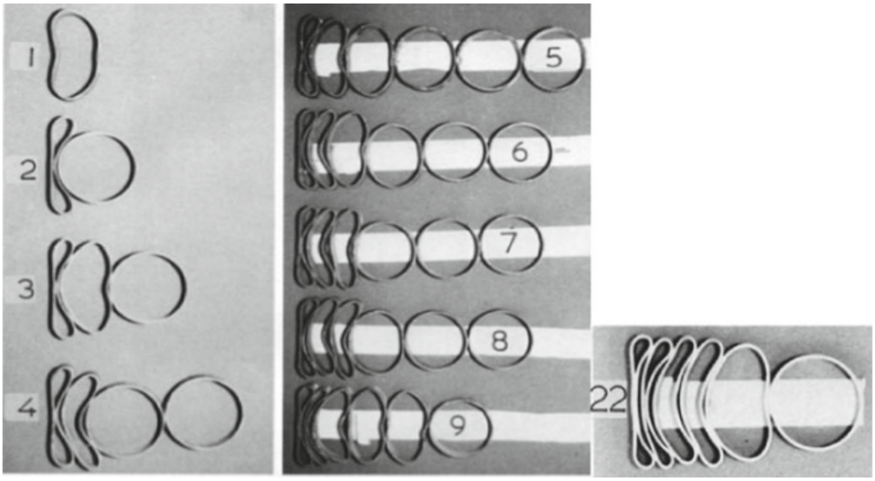


Fig. 1 Deformation in various ring systems from [6], with impact end to the left and free end to the right. 1–4 maintain similar impact velocity, varying the number of rings. 5–9 have increasing velocities. 22 is brass rather than mild steel and features similar impact velocity to 7

with those gathered experimentally, although it was noted that EST wasn't perfect for replicating the collapse at the hinges, and a different or developed method would be necessary to provide an optimised model.

One example of a study examining loaded rings using numerical analysis is Liu et al. [11]. This explores the stress wave propagation through ring systems, using both experimental and numerical methods. Using a split Hopkinson pressure bar test, the experimental section looked at single ring systems subjected to pulse loading, with a transmitter bar also acting as a fixed end for the systems. As in [7], the primary material was mild steel, although the rings were much thicker. Furthermore, the velocity was far lower, limited to 7 m/s to prevent any plastic deformation in the experimental set-up.

The numerical simulation also performed using a 3-dimensional analysis for better accuracy. A quarter-model was used in these simulations, as symmetry allowed the authors to justify a simplified model. One finding was that the position and order of maximum deformation in a ring system was affected by the elastic stress waves' reflection and superposition. Furthermore, this study helps to establish a basic methodology to be developed in a further study [12] while standing in its own right as an example of how rings subjected to loading can be analysed through numerical analysis.

2 FE Analysis

2.1 Numerical Analysis

A two-dimensional analysis approach was chosen, the advantage being that a much less complex mesh was required, thus greatly reducing computing effort. As the load was applied without any component in the z-axis (representing depth) and due to the ring's geometry being constant through this axis, no variation in stress or strain would be expected in a 3rd dimension in a model anyway, unless it was introduced by an uneven mesh. Therefore, it was decided that the lack of 3rd dimensional analysis wouldn't affect the accuracy of the results obtained.

A half-model was used, as under ideal circumstances a ring with perfect material properties and setup would deform symmetrically if impacted by a flat, rigid object confined to moving along a single axis. Using a full model would have introduced the possibility of a non-symmetric mesh affecting the mode of deformation across the symmetric axis. In the x–y plane this half-model appears similar to the model used in [11] and [12].

A dynamic analysis was necessary for the impact loading. For this, Abaqus (Simulia, US) was used, as it allows explicit, dynamic solutions and is ideal for contact. The model consisted of three parts, making it relatively simple. The ring was sandwiched between two rigid walls; one an impact body, one a base anvil. The ring used shell elements with a thickness assigned. The base anvil had no material properties, merely acting as the fixed, distal end. The rigid wall could easily be defined, and it was considered that its material properties does not need to be same as the ring, so any inertial properties could be applied.

Mild steel was chosen to model the ring, and appropriate properties assigned. These initially included the most important properties of density and the elastic data, including the Young's Modulus and Poisson's ratio. From this, the plastic properties were added to improve the model's accuracy, with the yield stresses and corresponding plastic strains used. The plastic region was defined as a simple linear relation. This was not modelled on a specific grade of steel- it was an approximation of the material. Considering this, the model could be argued to lack accuracy, but the intention is simply to give a general look at mild steel, so the approximate values used were considered adequate. No gravitational settings were applied, so this can be viewed as comparable to the sledge experiments in [5, 7], especially as being two dimensional restricts z-axis deformation.

As the velocity would be constant and the simulation would therefore not come to an end with the complete deceleration of the impact mass (as per [5, 7]), an appropriate time step would be applied to each model. The period that needed examining was the entirety of the possible crushing, until each ring was flattened but before direct compression of the material through the wall thickness. Therefore, the step was dependent on the specific ring dimensions and impact velocity, to allow full crushing.

The boundary conditions were applied as loads, the most significant being the loading of the large mass, applied as a velocity. To keep the half-model accurate, the displacement of all surfaces along the symmetrical axis was fixed in the x-direction by applying a velocity of 0 m/s and angular velocity of 0 s^{-1}. This prevented the model behaving as an independent half-ring and constrained it to deform in the same manner as a complete ring. Also, the base wire was completely fixed to act as a true fixed end. Meshing needed applying to all parts in the model, although the significance of the mesh in the two wires was negligible. The ring dimensions were the same as those used predominantly in [5, 7], largely to allow some initial comparison.

A quadrilateral mesh was used, as it was deemed to be best. To find the ideal number of elements, convergence was studied. Starting with an arbitrary number of elements around the circumference, an ideal number of elements through the wall thickness could be found. With this number of elements chosen, the number of circumference elements was optimized. The final mesh had 6 elements through the wall thickness and 540 around the inner and outer circumferences each.

2.2 Numerical Procedure for Ring Simulation

Initially, a range of different thickness-radius ratios and impact velocities was tested on single rings. The ranges of these parameters were 5 to 50 m/s impact velocity and a thickness-radius ratio of 1–10% (calculated as wall thickness T divided by outer radius r). Both looked at 10 even increments through their respective ranges.

The results required included a test condition to examine the full range of velocities on a fixed thickness. The thickness ratio (T/r) for this was set in the middle of the range at 5%, with the outer diameter measuring 50 mm. This diameter was similar, and so also comparable, to the size of rings used by Reid et al. It was also kept constant for all other tests, with different inner dimensions used to vary the thickness ratio when required. The results obtained from all experiments were the principal strains at each time step, as well as the reaction forces (*RF*).

The second series of tests involved varying the thickness-radius ratios on fixed velocities. The ratio was set for every value between 1 and 10%, and the three different velocities applied to each were 5, 25 and 50 m/s. This allowed comparison at certain velocities across the full range of thicknesses. Once a simulation was run, the results such as strain values and reaction forces were exported.

2.3 Numerical Procedure for Hollow Sphere Simulation

Despite the significant change to the geometry of the shape being modelled, the basic model set-up appeared much the same. This is due to an axisymmetric approach adopted, meaning that a 2D analysis in the same manner as the rings could be performed on a cross section of the sphere. Importantly, this meant re-constructing each part as axisymmetric rather than two-dimensional shell elements.

The same mass and inertia settings were applied as for the rings. The same material was used, with the same properties, as this had the obvious benefit of allowing direct comparison.

Compared to the procedure used for the rings, a reduced sample size was used. This involved a slightly reduced range of thicknesses, ranging from 3 to 9%, and only at 2% increments. Also, although same range of velocities was used, only 5, 15, 25, 35 and 50 m/s impact velocities were used on the 5% thickness ring, with only 25 and 50 m/s were applied to all others. Otherwise, the general procedure was kept much the same for the spheres to how it had been for rings, with the same results taken in the same manner.

3 Results

3.1 Ring Model

This section shows select comparisons of rings simulated under different conditions in the parametric tests. Here, comparisons are made of a ring at a constant thickness being crushed at different velocities. A visual analysis is presented in Table 1, which shows select velocities applied to the 5% thickness ring.

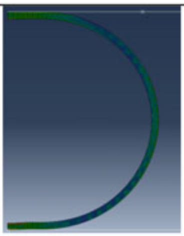
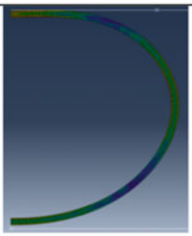
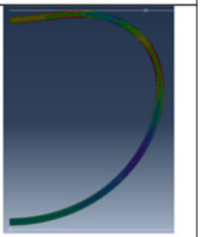
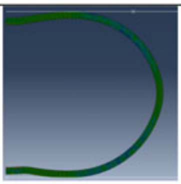
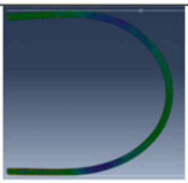
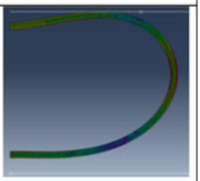
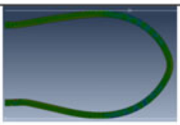
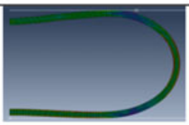
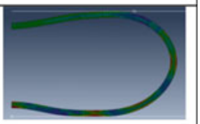
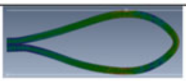
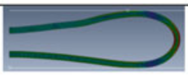
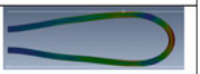
Figures 2, 3 and 4 show specifically the 5% thickness ring, as this was tested with the complete range of velocities. All these velocities are displayed in the graphs, which show various features, these respectively being the maximum strain value, the maximum strain location, and two graphs showing the loading, all plotted against displacement.

Further results were collated and compared for rings of different thicknesses being crushed at a constant velocity, namely 25 m/s.

3.2 Hollow Spheres Model

This section shows select comparisons of spheres simulated under different conditions in the parametric tests. As with the rings before, comparisons were made of a sphere at a constant wall thickness being crushed at different velocities. Here however, spheres of different wall thicknesses being crushed at a constant velocity (25 m/s) are compared. Table 2 shows a selection of these from the range of thicknesses, while Figs. 5 and 6 show the graphical data from the full range of thicknesses.

Table 1 A visual comparison, showing rings of the same thickness (5%) being crushed at different velocities. The displacements shown are also equivalent to crush percentages of 20, 40, 60 and 80%

Crush (%)	Velocity (load rate)		
	5(m/s)	25(m/s)	50(m/s)
20			
40			
60			
80			

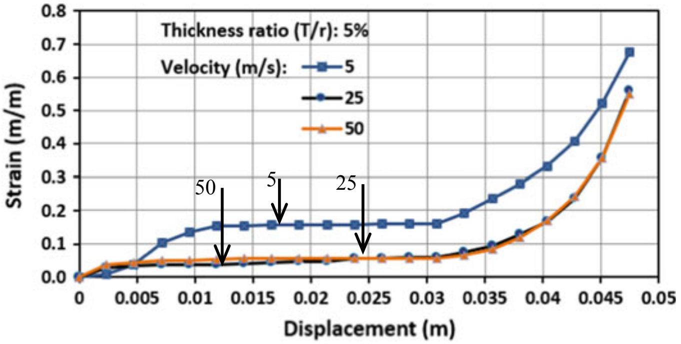


Fig. 2 Graph showing maximum strain values in a ring of 5% thickness being deformed at different rates, plotted against displacement

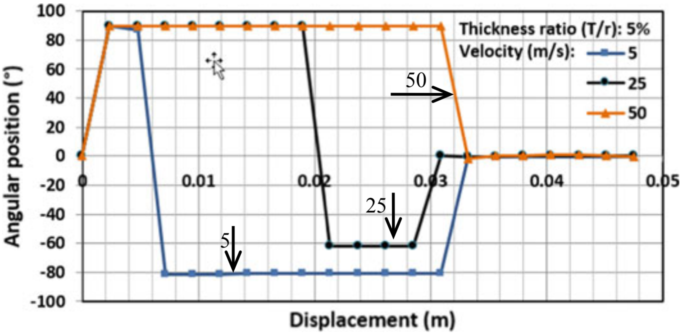


Fig. 3 Graph showing angular position of the ring (with 5% thickness) being deformed at different rates, plotted against displacement

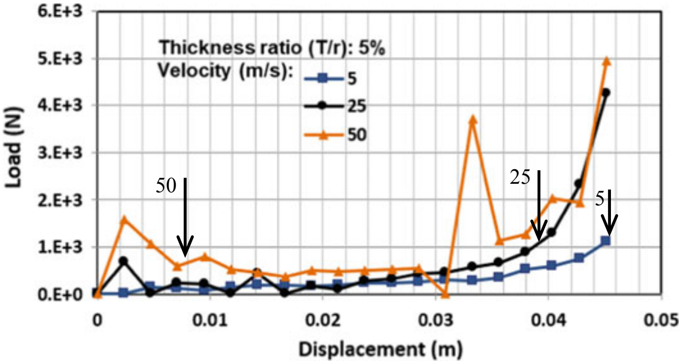


Fig. 4 Load displacement graph for a ring of 5% thickness being deformed at different rates

3.3 *Rings Versus Spheres Comparisons*

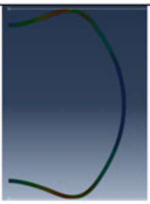
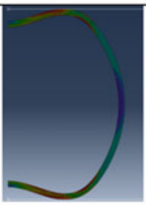
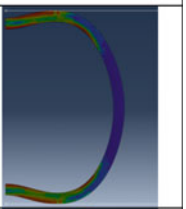
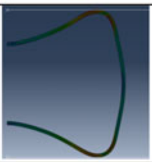
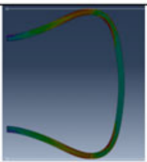
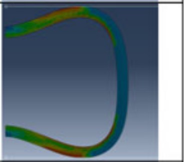
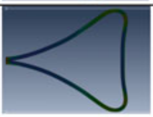
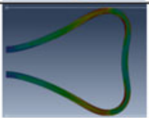
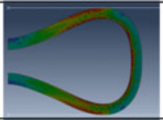
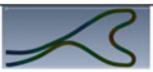
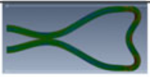
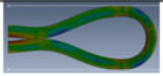
This section shows results comparing the deformation process of rings and hollow spheres. Figures 7 and 8 show graphical data of a ring and a sphere of equal wall thickness ratios and with the same loadings (5% and 25 m/s), for strain and load respectively.

4 Discussion

4.1 *Rings Modes of Deformation and Comparisons*

One of the clearest observations that can be made from the visual data of the rings is that many experience a ‘bounce’, whereby the lower region lifts away from the base

Table 2 A visual comparison, showing spheres of different thicknesses being crushed at the same velocities (25 m/s)

Crush (%)	Thickness ratio (*T*/*r*)		
	3(%)	5(%)	9(%)
20			
40			
60			
80			

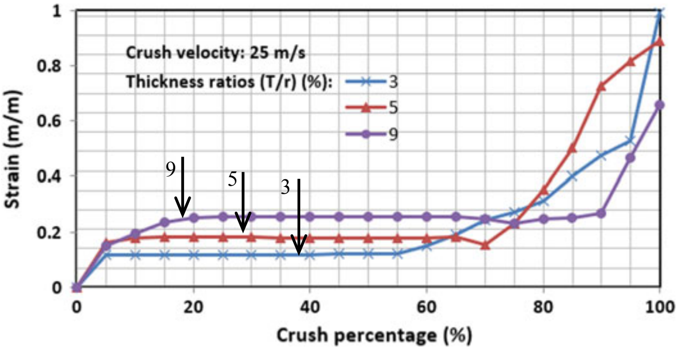


Fig. 5 Graph showing maximum strain values in spheres of various wall thicknesses being deformed at a rate of 25 *m/s*, plotted against crush percentage

wire for a period of time. This leaves none of the ring in contact with the fixed end, meaning that the ring is being freely pushed by the impact wire with no resistance other than the ring's own inertia. It appears that the bounce is a function of the impact velocity and ring thickness, as its onset and duration seems dependent on both. For most ring thicknesses there is little or no bounce at the 5 m/s impact velocity, whereas 25 m/s results in a moderate bounce and 50 m/s a large one. This is large in that the

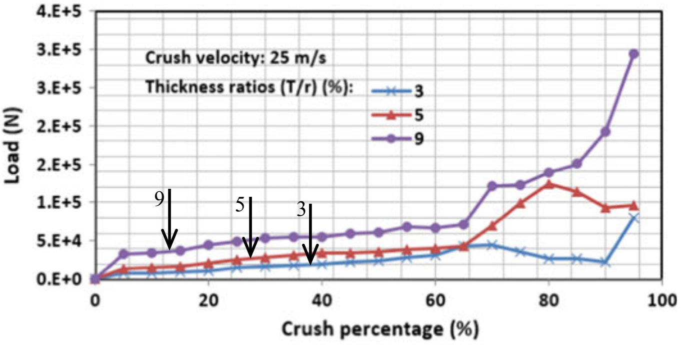


Fig. 6 Load displacement graph for a ring of various thickness being deformed at rate of 25 m/s

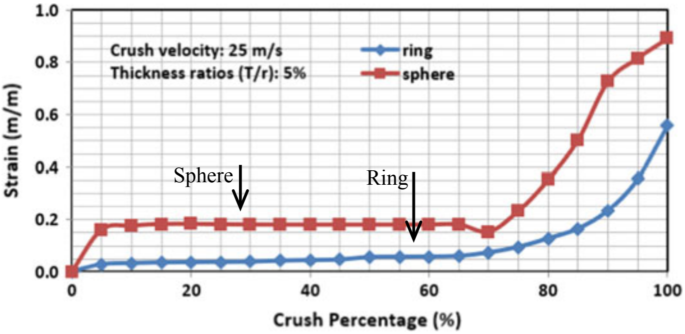


Fig. 7 Graph comparing the maximum strain values for a ring and hollow sphere

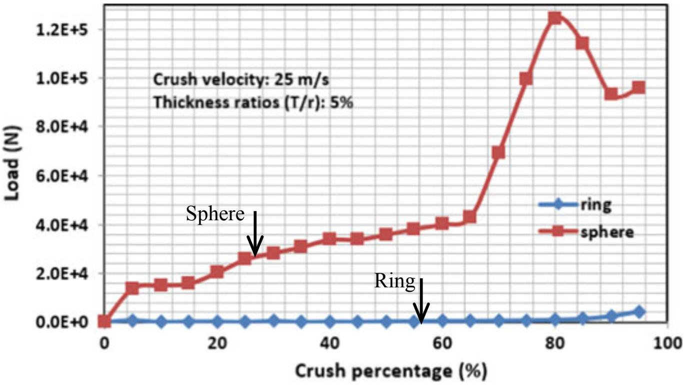


Fig. 8 Graph comparing the load crush displacement curves for a ring and hollow sphere

bottom of the ring travels a further from the fixed end and the bounce lasts longer. The wall thickness seems to impact the bounce, reducing its severity as the thickness is increased. The thinnest rings are particularly reactive, as even at 5 m/s the 1% thickness ring clearly bounces at 20% crushing.

As shown in Fig. 2, when velocity is increased, strain reaches a slightly higher initial value. This strain immediately stabilises, with higher velocity cases remaining most stable while others gradually increase, here, higher velocity cases a deeper curve. The same trend is familiar to other ring thickness ratios too, although thicker rings record a higher strain value for the stable region and have shallower curves than those of thinner rings.

The angular positions of these maximum strain values are more complicated, with little in the way of a clear trend. What is clear is that approximately the final third of the deformation process results in maximum strain occurring at 0°, this being the hinge located at the side and level with the ring's centre. The majority of other locations are at or close to the very top and bottom of the ring. This relates to the initial compression of the top-most section and the subsequent dip that occurs both at the impact and distal ends, resulting in the hinges at each location.

4.2 Hollow Spheres Mode of Deformation

The typical process of deformation for a sphere features the upper-most central section immediately dipping, creating a clear hinge which is in contact with the impact wire. This dip continues throughout, meaning the downwards pointing section grows in length as the hinge moves further from the centre of the model. The distal section acts in the same way, slightly after the upper section, so that, similar to rings, a bow shape is formed. But unlike the rings, the spheres have the two distinct hinges in contact with either wire, rather than one centrally at the side like for rings. This causes the side portion of the sphere—almost becoming tangential to the two wires—to stay relatively smooth (almost un-deformed), and as the hinges move away from the centre in the x-direction, this side section straightens (Table 1 at 40–60% crush).

This means that once the hinges have finished travelling outwards along the wire, there is a considerable amount of material (the hinge) directly above and below the now fairly straight side section. Here, the spheres behave differently depending on their wall thickness. With the larger wall thicknesses, the hinges aren't as tight (have a larger radius), meaning that the bulk of the hinges are, for the most part, still inside the side-section. Therefore, this section hinges in the middle in an outwards direction. As all three hinges are now in the same direction, the upper and lower ones open and merge into the new side one, so the sphere begins to fold in two. However, with thinner walls there are tighter radius hinges, meaning there's more material outside of the side-section. This causes the collapse of the side to hinge inwards, meaning the sphere folds in four instead, as in Table 2. Between 5 and 7% wall thickness ratio is the critical value which determines the deformation process a sphere will take. All this is illustrated in Table 2.

Unlike the rings, no complicated bounce process occurs. The strain values for different velocities follow a distinctive pattern whereby they rise briefly, then stabilise. Mostly, all reach a similar strain value, although the higher the velocity the earlier (in terms of displacement) they reach it. A large stable period follows whereby there is virtually no strain change (with the exception of 5 m/s loading, which follows similar pattern, but with a higher strain value that takes longer to reach before stabilising). There is then a distinct point at 0.029 m of displacement where all spheres sharply increase in strain value. According to the visual data, this occurs soon after the central hinge (pointing inwards for this wall thickness, leading to folding in 4) has started to develop. This is confirmed by Figs. 2, 3 and 4 showing that at the same displacement the maximum strain is found to move closer to the 0° value. Here, the maximum strain position tends to oscillate around positive and negative 20°, trending closer to 0° before settling there. As appears to be typical for rings and hollow spheres, the 5 m/s velocity does not follow this trend. This could be related to findings in [12, 13] which identified the point when maximum deformation would occur at the impact end, rather than the distal end as being 8.8 m/s for rings of a 50 mm diameter in that given system. When the velocity is unchanged and the wall thickness varied (as shown in Fig. 2), the general strain pattern remains the same.

In terms of loading, the spheres appear to show significantly less oscillating than the rings. At 5% wall thickness they similarly reach an initial peak and do not then experience a low loading period. Instead, there's a brief period where strain levels out slightly. The extent to which this happens seems to be a function of velocity, as the lower velocities maintain a steeper gradient, and faster impacts actually create a slight negative gradient. The load then increases linearly through the middle of the displacement range, with higher velocities returning slightly steeper gradients.

4.3 Rings Versus Spheres Comparisons

The distinct difference in deformation processes is further corroborated by the much sharper transition that a sphere has from a period of stable strain to one of rapidly increasing strain, as can be seen in Fig. 7. Finally, Fig. 8 shows how much stronger the sphere is despite sharing a cross-section, as it requires significantly larger loading to deform at all, let alone be completely crushed. However, it is noteworthy that this is only a comparison between wall thicknesses of 5%, and that the very different process of spheres thicker than this makes conclusive comparisons difficult.

References [14–17] investigate a numerical modeling of pimples or plates under heat conditions such as welding process. The current study did not consider the effect of weld on specimens and assumed the specimens made of single manufacturing process such as forming of thine plates. If any welding or other attachment is used on specimens then the material and boundary conditions with numerical modelling should be considered. Furthermore, the generated heat during the deformation and at the time of self-contact of the either ring or sphere materials has not been affected in numerical modelling and left it for future study.

Creating bilayer specimens using metallic and nonmetallic layers [18] was initial interest of the project but could not be completed due to time scope of the project. Finally, study on fatigue behavior [19] of the specimen under an impact and crash loading would be beneficial for a proper design in this field which would be the interest for next study.

5 Conclusion and Future Work

This paper is a revised and extended version of [20]. The general process of deformation of rings has been analysed, showing that a bow shape is typically formed by the dipping of the central impact end of the ring and eventual rising of the central distal end. It has also been found that a bounce causes the distal end to lift earlier. Although the reason for this is not certain, it seems to be a result of shock and a function of the ring's velocity and thickness. These factors are also responsible for the loading and strain characteristics that a particular ring will exhibit. The great oscillation in loading of rings at high velocity is also apparent, and this in turn could be related to the mechanism causing the ring to bounce. Spheres have also been analysed and their main deformation patterns recorded, again including strain, and loading characteristics. This has also allowed comparisons to rings that highlight the hollow sphere's much improved resistive qualities, as well as the very different ways that spheres deform dependent on their wall thickness.

References

1. V.P.W. Shim, R. Lan, Y.B. Guo, L.M. Yang, Elastic wave propagation in cellular systems-Experiments on single rings and ring systems. Int. J. Impact Eng **34**(10), 1565–1584 (2007)
2. T.V. Karman, P. Duwez, The propagation of plastic deformation in solids. J. Appl. Phys. **21**, 987–994 (1950)
3. S.E. Jones, P.J. Maudlin, J.C. Foster Jr., An engineering analysis of plastic wave propagation in the Taylor test. Int. J. Impact Eng **19**(2), 95–106 (1997)
4. G.R. Johnson, T.J. Holmquist, Evaluation of cylinder-impact test data for constitutive model constants. J. Appl. Phys. **64**(8), 3901–3910 (1988)
5. S.R. Reid, T. Yella Reddy, Experimental investigation of inertia effects in one-dimensional metal ring systems subjected to end impact—I. Fixed-ended systems. Int. J. Impact Eng. **1**(1), 85–106 (1983)
6. S.R. Reid, W.W. Bell, R.A. Barr, Structural plastic shock model for one-dimensional ring systems. Int. J. Impact Eng **1**(2), 175–191 (1983)
7. T.Y. Reddy, S.R. Reid, R. Barr, Experimental investigation of inertia effects in one-dimensional metal ring systems subjected to end impact—II. Free-ended systems. Int. J. Impact Eng. **11**(4), 463–480 (1991)
8. S.R. Reid, W.W. Bell, Response of 1-D metal ring systems to end impact. Inst. Phys. Conf. Ser. **70**, 471–478 (1984)
9. T.Y. Reddy, S.R. Reid, J.F. Carney III, J.R. Veillette, Crushing analysis of braced metal rings using the equivalent structure technique. Int. J. Mech. Sci. **29**(9), 655–668 (1987)

10. W. Merchant, On equivalent structures. Int. J. Mech. Sci. **7**(9), 613–619 (1965)
11. K. Liu, K. Zhao, Z. Gao, T.X. Yu, Dynamic behaviour of ring systems subjected to pulse loading. Int. J. Impact Eng **31**(10), 1209–1222 (2005)
12. J. Shen, K. Liu, G. Lu, Impact behaviour of a multi-body system with energy dissipation. Int. J. Crashworthiness **13**(5), 499–510 (2008)
13. M. Chizari, M.L. Barrett, Single and double plate impact welding: experimental and numerical simulation. Comput. Mater. Sci. **46**, 828–833 (2009)
14. B. Kogo, B. Wang, L. Wrobel, M. Chizari, Authentication in welded clad plate with similar material and thickness, in *Proceedings of the Eighth International Conference on Advances in Civil, Structural and Mechanical Engineering*—CSM 2019, Birmingham, UK (2019). https://doi.org/10.15224/978-1-63248-170-2-01
15. W. Guo, L. Zhang, C. Xu, R. Chai, Z. Gao, B. Kogo, M. Chizari, C. Zhang, B. Wang, Study on the wear resistance of laser cladding iron-base alloy by heat treatment. Mater. Res. Express 6(2) (2019) Article number 026572. 10.1088/2053-1591/aaf251
16. B. Kogo, B. Wang, L. Wrobel, M. Chizari, Experimental and numerical simulation of girth welded joints of dissimilar metals in clad pipes. Int. J. Offshore Polar Eng. **28**(4), 380–386 (2018). https://doi.org/10.17736/ijope.2018.oa22. ISSN 1053-5381
17. B. Kogo, B. Wang, L. Wrobel, M. Chizari, Microstructural analysis of a girth welded subsea pipe. Eng. Lett. **26**(1), EL_26_1_23 (2018). http://www.engineeringletters.com/issues_v26/issue_1/EL_26_1_23.pdf. ISSN: 1816-0948 (online version); 1816-093X (print version)
18. A. Khodadadi, G. Liaghat, D. Shahgholian-Ghahfarokhi, M. Chizari, B. Wang, Numerical and experimental investigation of impact on bilayer aluminum-rubber composite plate. Thin–Walled Struct. **149**, 106673 (2020). https://doi.org/10.1016/j.tws.2020.106673
19. N. Amiri, G.H. Farrahi, K.R. Kashyzadeh, M. Chizari, Applications of ultrasonic testing and machine learning methods to predict the static & fatigue behavior of spot-welded joints. J. Manuf. Process. **52**, 26-342020 (2020). https://doi.org/10.1016/j.jmapro.2020.01.047
20. M. Farr, B. Kogo, B. Wang, M. Chizari, Experimental and Numerical Modelling of Metal Rings Subjected to Fixed-End Crush Loading, and Comparisons with Hollow Spheres, in *Proceedings of the World Congress on Engineering 2019*, WCE 2019, July 3–5, London, UK (2019)

Investigation of Anterior Cruciate Ligament of the Knee with Relevance to Surgical Reconstruction—A Planar Mathematical Analysis

Ahmed Imran

Abstract Anterior Cruciate Ligament of the knee is injured quite often while performing strenuous activities like in sports. Due to poor healing characteristics, surgical reconstruction of the ligament is used to restore the joint function. However, many significant percentage of the patients are unable to return to their pre-injury levels of activity. In addition, more complications of the joint can result in repeated surgeries. Several clinical and experimental reports suggest the need for further investigations in order to gain insight in the behavior of the reconstructed ligament and related outcome. The present study used mathematical modelling to simulate the knee function in the sagittal plane when the ligaments are intact. The ligaments were separated into fiber bundles similar to those reported in the literature. Such simulations could be used to investigate effects of different tunnel positions during the ligament reconstruction. Knee motion during 0°–120° flexion and an anterior laxity test at different joint positions were simulated. The outcome of the model simulations suggest that anterior fibers of the ligament contribute significantly throughout the knee flexion in resisting anterior forces on the tibia. In comparison, the posterior fibers contribute in near extremes of motion only. The results of model simulations corroborated with experimental observations from literature and have clinical relevance.

Keywords ACL injuries · ACL mechanics · Double bundle ACL reconstruction · Femoral tunnel in ACL reconstruction · Knee joint mechanics · Reconstruction of ACL

1 Introduction

Anterior Cruciate Ligament (ACL) of the knee is injured quite often while performing strenuous activities like in sports. This is more common in young athletes. Treatment of such injuries often involves surgical reconstruction of the ligament with the goal

A. Imran (✉)
College of Engineering and Information Technology, Ajman University, PO Box 346, Ajman, UAE
e-mail: a.imran@ajman.ac.ae; ai_imran@yahoo.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_4

of restoring stability and kinematics of the joint. Another goal is to restrict further long-term risks of complications like osteoarthritic changes, chondral or meniscal damage that often occur subsequent to the ligament injury [1–5]. Clinical reports on long-term outcomes of treatment give somewhat satisfactory results, however, a significant number of patients, about one-third, still remain problematic and report complications with difficulty in returning to their earlier level of activity [1, 5]. The reported outcome shows a wide variety of outcomes. For example, one study recorded less than 50% cases of athletes who were treated with reconstruction returned to their activities of pre-injury level [5]. In contrast, another clinical report recorded that 94% of patients from the surgery even after five years of follow-up continued to report knee instability [6]. Such observations suggest that there is a need for further investigations in order to gain insight in the behavior of the reconstructed ligament and related outcome.

The two cruciate ligaments, namely, anterior cruciate ligament or ACL and posterior cruciate ligaments or PCL are considered to be the main stabilizers of the knee joint in the sagittal plane [7–10]. ACL provides the main restrain to anterior movement of the lower tibial bone relative to the upper femoral bone. PCL provides the main restrain to posterior movement of the tibia relative to the femur.

Anatomical and histological studies of the ACL have classified the ligament into two distinct functional bundles of fibers with separate areas of clearly marked insertions on the femur and tibia [11–17]. These observations follow two approaches for surgical reconstruction—one treating the whole ligament as a single bundle while the other treating the ligaments as a combination of two distinct bundles [11–17].

Recent surgical approaches of reconstruction have given importance to the two bundles of fibers and have accordingly tried to restore the ligament with its original anatomy. However, surgical reconstruction with two bundles has also sometimes resulted in unexpected outcome where the joint produced abnormal instability and kinematics [2–5].

A clinical study of 90 patients with a 10-year follow-up showed fewer graft failures [2]. About one third of this group of patients had double bundle reconstruction of the ACL. The study also reported that 66% patients developed osteoarthritis of the knee with most sever changes in the patients who had the longest delay from the primary injury to ACL reconstruction and in the patients who underwent partial meniscal resection at the time of ACL reconstruction [2]. Other clinical studies further suggest that exact attachment of femoral site of the reconstructed graft critically influences the surgical outcome [3, 4, 14]. Steckel et al. [16] used radiographs to evaluate anterormedial (AM) and posterolateral (PL) bundles of the ACL. The researchers observed that beyond 90° flexion, centers of these fiber bundles become more horizontal. The investigators, hence, suggested that the degree of knee flexion should be taken into account for femoral tunnel placement and for describing tunnel positioning.

Kawaguchi et al. [14] analyzed the role of ACL fibers in resisting tibial displacement in cadaver knees. They reported major role for the AM and PL bundles of ACL in resisting anterior tibial displacement relative to femur.

In addition, other studies analyzed strains developed in the fibers of the ACL. Observations from such studies also support distinct role for the ACL fibers. Again, the AM bundle was found to be the primary restraint for anterior translation of tibia relative to femur. Further, the PL bundle bundle was found to provide contributions near the extremes of joint motion in the sagittal plane [17–19].

The aim of the present work is to study forces in distinct bundles of the ACL during a simulated laxity test which is similar to an anterior drawer test conducted at a fixed position of the joint. For this purpose, a sagittal plane mathematical model of the knee is used. The model ligaments were formed with bundles of elastic fibers having distinct insertions on their respective bones. Joint motion and the laxity test are simulated at several flexion positions of the joint.

2 Methods

Quantitative analysis of the ACL function required a model of the knee to simulate the joint kinematics and a laxity test at several flexion positions. The joint kinematics in the sagittal plant was simulated mathematically and the effects of laxity test similar to anterior drawer test were superimposed at selected positions.

The model parameters comprised anatomical shapes of tibial and femoral bones, shapes and positions of articular surfaces, points of attachment of the joint ligaments and material properties of ligaments. The model parameters were taken from anatomical studies reported in the literature [20–27].

The knee mechanics is described with four major ligaments—two cruciate ligaments and two collateral ligaments. A representation of non-linear elastic fibers was used for each ligament. The fibers resisted stretching and remained slack under compression. Further, consistent with the anatomical literature, the collateral ligaments had insignificant contribution in the sagittal plane mechanics of the joint.

2.1 Kinematics of the Knee During Flexion

The knee joint motion in the sagittal plane was simulated for 0–120° flexion. The knee kinematics in the absence of external loads or muscle action was defined as passive motion that required selected fibers in the cruciate ligaments to maintain isometric lengths as the bones rotated and moved relative to each other [25]. This modelled the joint kinematics in the unloaded state [26]. Resulting from relative rotations and translations of the bones, insertions of the ligaments on the respective bones also changed their relative positions. The mathematical model of the knee with intact ligaments and resulting joint kinematics was developed based on previously reported studies [20–28].

2.2 Simulated Anterior Laxity Test

During a laxity test like the Drawer test at a fixed joint position, the tibia is pulled anterior to the femur and resulting translation is recorded corresponding to a known applied force [29–32]. This effect was simulated by superimposing additional relative translation at insertion points of the ligament fibers. Such tests are conducted to estimate integrity of the ligaments [29].

During the model simulation of the laxity test, A flexing moment and an anterior force were applied on the tibia. As a result, the tibia translated and stretched the ACL fibers until equilibrium. In the model, joint angle and magnitude of the anterior force could be selected and resulting anterior translation of tibia (ATT) calculated. In addition, the model also allowed application of a predetermined value for ATT and calculation of associated forces in different fiber bundles of the ACL.

2.3 Distribution of Forces in Different Areas of the Ligament Attachments

The ACL was modelled as a combination of two bundles of fibers as reported in the literature. The anterior bundle represented the anteromedial fibers and the posterior bundle represented the posterolateral bundles of the ACL [7–9, 11–14]. Fibers in each bundle sequentially stretched or slackened during motion or due to the tibial translation relative to femur.

In the model simulation of the laxity test, the tibia was translated 8 mm anterior to its position achieved at each flexion angle during motion. As a result of the applied external force, the tibia translated anterior or posterior to its resting position recruiting ligament fibers progressively until the external force was balanced by the forces generated in the stretched segments of the involved ligaments.

3 Results and Analysis

3.1 Simulated Anterior Laxity Test

Table 1 gives a comparison of model calculations with experimental measurements. ATT calculations and measurements correspond to laxity test with 90 N anterior force on the tibia while the joint angle remained fixed at selected positions.

The experimental measurements with mean values and standard deviation were reported by Kondo et al. [15]. The model used a similar simulation with 90 N force.

The results of simulation and experimental show similar trends. With the flexion angle increasing from 0° to 110°, ATT first increased till about the mid range and then

Table 1 ATT calculations from the model knee are compared with in vitro measurements on 14 cadaver knees resulting from anterior laxity test with 90 N force [15]

Flexion angle (°)	ATT (mm)	
	Model calculations	Experiment [15] Mean (Std. Dev.)
0	2.6	1.8 (1.3)
30	5	4.0 (1.3)
60	5	3.8 (2.2)
90	4	3.0 (1.4)
110	3.7	2.9 (1.7)

decreased in high flexion. The values calculated at each position remained within the range of those measured.

Near extension, the fibers of ACL are less slack and more ready to stretch but oriented more perpendicular to the anterior direction resulting in less ATT when 90 N force is applied. In comparison, the fibers become slack in the mid flexion range and application of anterior forces translates them anteriorly, thus, stretching the ligament fibers sequentially that helps in developing progressive resistance in the ligament. Finally, in high flexion, the ligament fibers are oriented with reduced inclination to the anteroposterior direction, thus, providing a larger posterior component to balance the external anterior force.

3.2 *Distribution of Forces in Anterior and Posterior Bundles of the ACL*

Figure 1 shows posterior forces developed in the ligament fibers over 0°–120° flexion range. Contributions of the anterior and posterior bundles of the ACL given as a percentage of total anterior force applied to translate the tibia 8 mm anterior to the femur at each simulated flexion position.

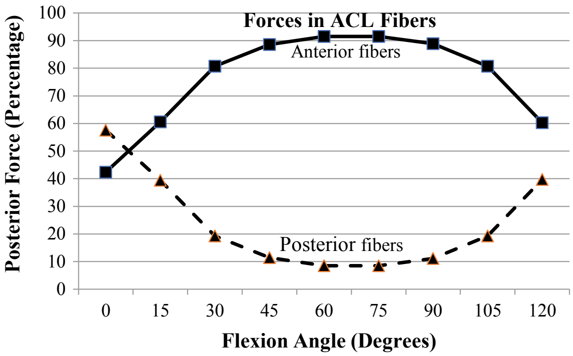


Fig. 1 Posterior force developed in the ligament is given over 0°–120° flexion. Forces in each of the anterior and posterior bundles of the ACL are given in terms of percentage of total anterior force on the tibia that resulted in 8 mm ATT simulation

At low flexion, both the fiber bundles contributed significantly in resisting the external force. With increase in flexion angle, contribution of the posterior bundle decreased sharply and that of the anterior bundle increased till about 45°. The anterior bundle resisted the external force fully in the mid range of the joint motion. Beyond 90°, the posterior fibers stretched again and provided their contribution.

Kawagachi et al. [14] also reported similar patterns in their in vitro experimental observations using 8 cadaver knees to analyze load bearing function of the ACL fibers in resisting tibial displacement. In the study, attachments of ACL fibers on the femur were cut sequentially and translating force were measured for 6 mm ATT.

The investigators reported that the anteromedial bundle of the ACL contributed 66–84% of the total resistance during 0–90° flexion. In comparison, the posterolateral bundle contributed nearly 16–9%. They also measured the torques required for internal or external rotation of the bones and found no significant effect of cutting of the ACL fibers, This later observation confirmed that the main function of the ACL is in resisting ATT.

4 Conclusions and Future Work

The model calculations showed general agreement with experiments on cadaver knees reported in separate studies in relation with patterns of anterior tibial translations and contributions of anteromedial and posterolateral bundles of ACL with distinct attachments on bones.

The analysis suggests that restraint provided by different fiber bundles in the ACL depends two important factors, namely, flexion position of the joint and overall magnitude of ATT. This is because of each of the flexion position and the ATT result in changed orientations of the ACL fibers.

Further, the joint flexion and the ATT also influenced relative contributions of the anterior and posterior bundles of the ligament.. The anterior bundle resisted anterior forces on the tibia at all positions, while the posterior bundle stretched mainly near extension or in high flexion positions.

Changes in attachment positions of the ligament fibers and resulting effects form further extension of this investigation.

5 Clinical Relevance

Position of attachment of the ACL graft during the ligament reconstruction critically determines the surgical outcome and ability of the patient to return to earlier levels of activity. The present analysis suggests that anteromedial bundles of the ACL in the intact knee contribute significantly towards resisting anterior forces on the tibia.

Further, the model as well as experimental observations suggest that the intact ACL permitted less translation of the bones near the extremes of motion than in the mid flexion range. This may indicated that rehabilitation exercises or activities involving extreme positions of the joint may prove to be more demanding on an anatomically reconstructed graft. The analysis has relevance to ACL-reconstruction and ACL rehabilitation.

Acknowledgements The author would like to thank College of Engineering and Information Technology at Ajman University, UAE for financial and research support provided during the development of this work.

References

1. K.L.P. Monaghan, H. Salem, K.E. Ross, E. Secrist, M.C. Ciccotti, F. Tjoumakaris, M.G. Ciccotti, K.B. Freedman, Long-term outcomes in anterior cruciate ligament reconstruction—a systematic review of patellar tendon versus hamstring autografts. Orthop. J. Sports Med. **5**(6), 1–9 (2017)
2. S. Jarvela, T. Kiekara, P. Suomalainen, T. Jarvela, Double-bundle versus single-bundle anterior cruciate ligament reconstruction: a prospective randomized study with 10-year results. Am. J. Sports Med. **45**(11), 2578–2585 (2017)
3. S. Irarrázaval, A. Masferrer-Pino, M. Ibañez, T.M.A. Shehata, M. Naharro, J.C. Monllau, Does anatomic single-bundle ACL reconstruction using hamstring autograft produce anterolateral meniscal root tearing? J. Exp. Orthop. **4**(17), 1–5 (2017)
4. V.S. Alfonso, J.C. Monllau, Acute anterior cruciate ligament tear surgery: repair vs reconstruction—when? in *The ACL–Deficient Knee: A Problem Solving Approach* (Springer, Berlin, 2013), pp. 203–310
5. C.L. Arden, N.F. Taylor, J.A. Feller, E.K. Webster, Return to sport outcome at 2 to 7 years after anterior cruciate ligament reconstruction surgery. Am. J. Sports Med. **40**, 41–48 (2012)
6. M.M. Murray, S.D. Martin, T.L. Marti, M. Spector, Histological changes in the human anterior cruciate ligament after rupture. J. Bone Joint Surg. **82**-A(10), 1387–1397 (2000)
7. S. Amiri, T. Derek, V. Cooke, U.P. Wyss, A multiple-bundle model to characterize the mechanical behavior of the cruciate ligaments. Knee **18**(1), 34–41 (2011)
8. M. Mommersteeg, L. Blankevoort, R. Huiskes, J. Kooloos, J. Kauer, Characterisation of the mechanical behavior of human knee ligaments: a numerical-experimental approach. J. Biomechanics **29**(2), 151–160 (1996)
9. A. Amis, G. Dawkins, Functional anatomy of the anterior cruciate ligament–fibre bundle actions related to ligament replacement and injuries. J. Bone Jt. Surg. (Br) **73-**B, 260–267 (1991)
10. D. Butler, M. Kay, D. Stouffer, Comparison of material properties in fascicle-bone units from human patellar tendon and knee ligaments. J. Biomech. **19**(6), 425–432 (1986)
11. M.R. Carmont, S. Scheffler, T. Spalding, J. Brown, P.M. Sutton, Anatomical single bundle anterior cruciate ligament reconstruction. Curr. Rev. Musculoskelet Med. **4**, 65–72 (2011)
12. P.B. Jorge, D. Escudeiro, N.R. Severino, C. Santili, R.L. Cury, A.D. Junior, L.G.B. Guglielmetti, Positioning of the femoral tunnel in anterior cruciate ligament reconstruction: functional anatomical reconstruction. BMJ Open Sport Exerc. Med. **4**(1) (2018). https://bmjopensem.bmj.com/content/4/1/e000420. Last accessed on 15 May 2020
13. A. Imran, Force distribution in femoral attachment of anterior cruciate ligament during drawer test—a planar model analysis. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019, 3–5 July, 2019, London, U.K., pp. 418–421
14. Y. Kawaguchi, E. Kondo, R. Takeda, K. Akita, K. Yasuda, A.A. Amis, The role of fibers in the femoral attachment of the anterior cruciate ligament in resisting tibial displacement. J. Arthroscopic Relat. Surg. **31**(3), 435–444 (2015)

15. E. Kondo, A.M. Merican, K. Yasuda, A.A. Amis, Biomechanical analysis of knee laxity with isolated anteromedial or posterolateral bundle deficient anterior cruciate ligament. J. Arthroscopic Relat. Surg. **30**(3), 335–343 (2014)
16. H. Steckel, V. Musahl, F.H. Fu, The femoral insertions of the anteromedial and posterolateral bundles of the anterior cruciate ligament: a radiographic evaluation. Knee Surg. Sports Traumatol. Arthrosc. **18**, 52–55 (2010)
17. A. Imran, Relating knee laxity with strain in the anterior cruciate ligament. Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2017, 5–7 July, 2017, London, UK, pp. 1037–1042 (2017)
18. A. Imran, Analyzing anterior knee laxity with isolated fiber bundles of anterior cruciate ligament, Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2016, 29 June–1 July 2016, London, UK, pp. 869–872 (2016)
19. W. Petersen, T. Zantop, Anatomy of the anterior cruciate ligament with regard to its two bundles. Clin. Orthop. Relat. Res. **454**, 35–47 (2007)
20. V.B. Duthon, C. Barea, S. Abrassart, J.H. Fasel, D. Fritschy, J. Ménétrey, Anatomy of the anterior cruciate ligament. Knee Surg. Sports Traumatol. Arthrosc. **14**(3), 204–213 (2006)
21. A. Imran, J.J. O'Connor, Control of knee stability after acl injury or repair: interaction between hamstrings contraction and tibial translation. Clin. Biomech. **13**(3), 153–162 (1998)
22. A. Race, A. Amis, The mechanical properties of the two bundles of the human posterior cruciate ligament. J. Biomech. **27**(1), 13–24 (1994)
23. A.B. Zavatsky, The functional architecture of human knee ligaments. Ph.D. thesis, University of Oxford, 1993
24. A. Zavatsky, J. O'Connor, A model of human knee ligaments in the sagittal plane: Part 1. Response to passive flexion. J. Eng. Med. **206**(H), 125–134 (1992)
25. J.J. O'Connor, T.L. Shercliff, E. Biden, J.W. Goodfellow, The geometry of the knee in the sagittal plane. Proc. Inst. Mech. Eng. (Part H) J. Eng. Med. **203**, 223–233 (1989)
26. A. Imran, Modelling and simulation in orthopedic biomechanics-applications and limitations, in *Computational and Experimental Biomedical Sciences: Methods and Applications* ed. by J. M Tavares, R.M. Jorge (Springer, Berlin, 2015)
27. A. Watanabe, A. Kanamori, K. Ikeda, N. Ochiai, Histological evaluation and comparison of the anteromedial and posterolateral bundle of the human anterior cruciate ligament of the osteoarthritic knee joint. Knee **18**(1), 47–50 (2011)
28. T.W. Lu, J.J. O'Connor, Fiber recruitment and shape changes of knee ligaments during motion: as revealed by a computer graphics based model. Proc. Inst. Mech. Eng. (Part H) J. Eng. Med. **210**, 71–79 (1996)
29. J. Kupper, B. Loitz-Ramage, D. Corr, D. Hart, J. Ronsky, Measuring knee joint laxity: a review of applicable models and the need for new approaches to minimize variability. Clin. Biomech. **22**, 1–13 (2007)
30. A. Imran, Sagittal plane knee laxity after ligament retaining unconstrained arthroplasty: a mathematical analysis. J. Mech. Med. Biol. **12**(2), 1–11 (2012)
31. A. Imran, Influence of flexing load position on the loading of cruciate ligaments at the knee—a graphics-based analysis, in *Computational and Experimental Biomedical Sciences: Methods and Applications*, ed. by J.M Tavares, R.M. Jorge (Springer, Berlin, 2015)
32. A. Imran, Computer graphics based analysis of loading patterns in the anterior cruciate ligament of the human knee, in *Advances in Intelligent Systems and Computing*, ed. by K. Arai et al. (Springer Nature, Berlin, 2019), pp. 1–6. (https://doi.org/10.1007/978-3-030-01177-2_98)

Mathematics in Motion: A Model for the Ma Lin Ghost Serve

Hameez Mohammed, Rajesh Lakhan, and Donna M. G. Comissiong

Abstract In the sport of table tennis, the overall success of any given player is significantly correlated to the points gained on serves. Perfection of the serve is therefore an essential component of the training regime of all professional players. The Ma Lin ghost serve is a particular type of backspin short serve which is quite advantageous. In this paper, we propose a simple two-dimensional mathematical model to describe the behaviour of the ball during this short serve from initial toss to the third board bounce. For this purpose, we subdivide the entire serve into eight phases: four of which were points of contact between the ball and the racket or board, and four airborne trajectories. The first airborne phase is the toss which leads to the backspin serve which is the first contact phase. We account for the Magnus Lift generated by the spin throughout the next three airborne trajectories, and the effect of air resistance. We also consider velocity and rotational changes due to the respective coefficients of restitution and surface grip on the spinning ball as it bounces on both sides of the board to fulfill the requirements of a legal short serve and develop into a ghost serve.

Keywords Backspin serve · Ball trajectories · Ghost serve · Ma Lin · Mathematical model · Table tennis

H. Mohammed · R. Lakhan · D. M. G. Comissiong (✉)
Department of Mathematics and Statistics, The University of The West Indies, St Augustine Campus, St Augustine, Trinidad and Tobago
e-mail: Donna.Comissiong@sta.uwi.edu

H. Mohammed
e-mail: hameezmohammed@yahoo.com

R. Lakhan
e-mail: Rajesh.Lakhan@sta.uwi.edu

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_5

1 Introduction

The sport of table tennis has become increasingly competitive in recent years. While undergoing intense training regimes to hone their reflexes, professional players seek to expand their arsenal of attacking shots while simultaneously boosting their defensive plays. The psyche of a given competitor fluctuates according to the evolution of the match or the response of a given opponent. Individual levels of dexterity and the tendency or preference for various strokes contribute towards making every table tennis player unique. The success rate of the tactics employed by a player is dependent on the efficient implementation of different strokes with subtle variations in incident racket-ball angles, velocities and points of contact.

The overall performance of a player is significantly correlated to the points gained from serves [5]. The serve is initiated when the ball is tossed vertically upwards from rest in the open palm of the server's hand (without imparting spin). The ball rises at least 16 cm and then falls without making contact with anything, before being struck by the racket for the serve. This neutral beginning quickly morphs into a battle of wit, skill and technique. The initial strategy of attack—aimed at immediately or ultimately securing the point—should therefore commence at this stage. Our aim is to propose a mathematical model for a particular type of serve that, when executed well, would severely hamper an aggressive return by the receiver. For a given service to be legal, the ball must bounce once on the server's side of the table and subsequently at least once on the opponent's side. The serve is said to be short when the ball bounces more than once on the opponent's side.

It is a well-known fact that a ball incident with backspin on a horizontal surface has a tendency to bounce backwards [3]. In a backspin serve, the top of the ball spins toward the server through a horizontal axis parallel to the net. This backspin imparts an acceleration on the ball toward the server which causes the ball to slow down, and perhaps to reverse its direction (away from the opponent) during the serve. Players also contend that sufficient backspin on the ball can cause the ball to pitch sharply downward upon contact with the opponent's racket. This makes the successful return of a backspin serve difficult to achieve, and significantly decreases the possibility of the opponent countering with an attacking stroke.

The backspin serve attributed to the Olympic and four-time World champion Ma Lin is a short service that is commonly referred to as the Ghost Serve. It is produced by imparting a heavy (high revolution) backspin on the ball with minimal translational velocity. This allows the ball to have a much lower height when clearing the net in comparison to a serve with no backspin and to bounce twice on the receiver's side of the table unless it is returned in time. When enough backspin is imparted, after bouncing for the first time on the opponent's end, the ball reverses direction and spins towards the net, away from the opponent. This complex behaviour of the ball severely restricts the choice of possible returns from the opponent, which is to the server's advantage.

Our two-dimensional mathematical model for the Ma Lin ghost serve utilises the well-known kinematic equations of motion. We also incorporate linear and angular momentum conservation principles. In what follows, we present a step by step trajectory analysis for a given incident racket velocity and angle. We consider the effect of quadratic backspin, air resistance (drag) and Magnus lift on the ball. We utilise horizontal and vertical coefficients of restitution with a surface grip analysis for board-ball contact. Finite difference techniques are employed to solve the resulting equations numerically at each stage of motion in order to generate ball trajectories that resemble the Ma Lin ghost serve [6].

2 The Mathematical Model

The Ma Lin ghost serve for a table tennis ball is analysed up to the third board-ball collision point. Assuming that there is no sideways motion, we investigate the two-dimensional trajectory of the ball by partitioning the serve into separate phases. This subdivision divides the serve into four airborne and four contact phases, as depicted in Fig. 1.

The serve is initiated when the ball is tossed (without spin) vertically upwards from the open palm of one hand. The ball reaches a maximum height, which we refer to as point A. Our analysis of the serve begins when the ball falls from rest at the maximum height A. The first airborne phase therefore represents the vertical drop from the maximum height at A to the point B as indicated in Fig. 1. The first contact phase takes place at B when the server hits the ball with the racket to impart the necessary backspin. The second airborne phase BC is the trajectory of the ball after impact with the racket at B until the ball lands for the first time on the board at point C. The collision between the ball and the board at C is referred to as the second contact phase. It should be noted that for the serve to be legal, point C must be located

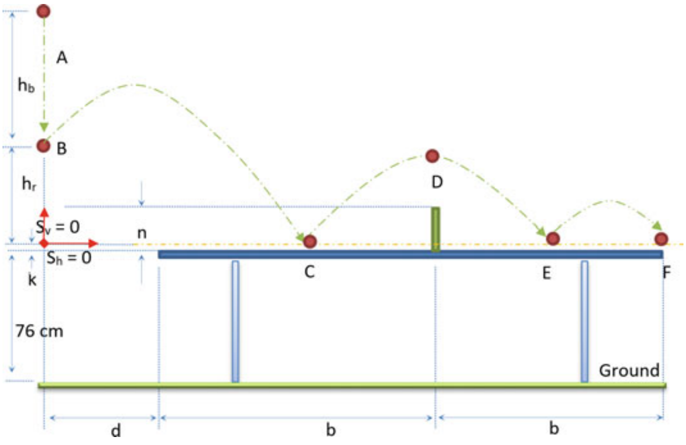


Fig. 1 A basic short serve with four airborne phases and four contact phases

on the server's side of the board. The ball subsequently rebounds from C and travels until it lands for a second time on the receiver's board at point E. The serve is said to be illegal if the second bounce at E does not take place on the receiver's end of the board. The third contact phase of the serve considers the impact that the bounce at position E has on the subsequent motion of the ball. During the third airborne phase CE, the ball must have cleared the table tennis net. The position of the ball when it is directly above the net is indicated by the point D. Phase EF investigates the rebound of the ball from E until it lands again at F.

The position of F is instrumental in rating the success of the serve. Point F must be located on the table tennis board for the serve to be classified as short. The Ma Lin ghost serve is a short serve where the spin on the ball reverses the direction of its trajectory after the second board-ball contact at E. It follows that if point F is located closer to the server than point E, the ball would have effectively reversed direction.

The individual style of each player is unique, and this is taken into account in our investigation. Key factors that may vary from player to player include the height of the ball toss, the serve position and the spin on the ball. The system also has fixed standards for regulated table tennis equipment. The radius of the ball k is $0.02\,\text{m}$, the upper surface of the board must be $0.76\,\text{m}$ above the ground, the length of half of the board b is $1.37\,\text{m}$, the height of the net is $0.1525\,\text{m}$ and the width of the board is $1.525\,\text{m}$ [10].

We adopt a two-dimensional Cartesian coordinate system for our analysis. The horizontal axis originates in the line of the ball toss AB and proceeds in a positive direction towards the receiver's end. The vertical axis commences one ball radius above the board's surface and takes the positive direction as vertically upwards. The values of velocities and displacements with respect to time are investigated via perpendicular horizontal and vertical axes. In generating the results of each phase, the final values of velocity and displacement in one phase are the initial values of the consecutive phase. The corresponding ball trajectories are calculated and displayed graphically.

2.1 Phase AB

We assume that the the ball is dropped vertically downwards from the maximum height A, to the point of contact with the racket B. Since the ball has no spin from the initial toss, no Magnus effect is considered during this phase of motion. We also assume that energy is conserved throughout this period. This allows us to calculate the horizontal and vertical components of velocity prior to impact with the racket in Eqs. 1 and 2.

2.2 Phase B

In the racket-ball collision phase, the racket moves from left to right with a velocity v_r . The racket strikes the ball at position B with an angle of elevation θ to the horizontal in a counterclockwise direction. The impact is at the base of the ball and this effects changes in its vertical, horizontal and rotational motions. For simplicity, we assume that the racket and ball coalesce, and that momentum is conserved. The initial velocities of phase B are the final velocities of the previous phase. Horizontal and vertical components of velocity are calculated directly after impact with the racket via Eqs. 3 and 4. We assume the ball leaves the racket with a backspin of N_{BC} revolutions per second (Table1).

Table 1 Nomenclature

Symbol	Units	Description
v_r	ms^{-1}	Velocity of serve
θ	degrees	Angle of elevation to horizontal
N_{BC}	revs. per sec	Backspin imparted by the racket
W	N	Weight of the ball
A_R	N	Force due to air resistance
M	N	Force due to magnus lift
SP	–	Spin parameter
C_D, C_L	–	Coeff. of drag, lift
e_{H_C}, e_{V_C}	–	Coeff. of horizontal, vertical restitution
K	m	Radius of the ball
$\omega 1_{BC}$, $\omega 2_{BC}$	rad. per sec	initial, final angular velocity BC
$u_{h_{BC}}$, $u_{v_{BC}}$	ms^{-1}	Initial horizontal, vertical velocity BC
$v_{h_{BC}}$, $v_{v_{BC}}$	ms^{-1}	Final horizontal, vertical velocity BC
F_{H_C}	N	Horizontal force at C
F_{R_C}	N	Frictional force at C
R_C	N	Reaction force at C
G	ms^{-2}	Acceleration due to gravity
h_b	m	Height of ball toss above board at A
h_r	m	Height of serve above board at B
m_b, m_r	kg	Mass of ball, racket
Re	–	Reynold's number
ρ_a	kgm^{-3}	Density of air
q_d	Nm^{-2}s^2	Proportional coeff. of drag
q_m	Nm^{-2}s^2	Proportional coeff. of Magnus lift

2.3 Phase BC

In flight, the ball is a light spherical object of weight W which loses a significant percentage of its energy due to air resistance A_R [7]. This air resistance or drag force acts on the airborne trajectory phases BC, CE and EF. As the Reynolds number is proportional to velocity, it follows that different serve velocities will correspond to a range of possible Reynolds numbers. The drag coefficient is approximately constant for the range of Reynolds numbers that we consider [1]. We may therefore conclude that the drag force is proportional to the square of the velocity, and this is reflected in Eq. 5.

The Ma-Lin serve requires a significant amount of backspin on the ball. This produces a variation of speed and pressure difference between the lower and upper surfaces of the ball. The effect of backspin is considered during the airborne phases BC, CE and EF. The pressure difference (Bernouilli's effect) generates a perpendicular force to motion usually referred to as the Magnus Lift—denoted by M—as seen in Fig. 2.

Rotation produces a factor known as the spin parameter SP which is defined as the ratio of the velocity of a peripheral point of the ball to the velocity of the centre of gravity of the ball [4]. We assume for this system that the spin parameter SP is 3.0 and that the racket imparts an angular velocity $\omega_{i_{BC}}$ on the ball. For this spin parameter, the coefficient of drag C_D is 0.73 and the coefficient of lift C_L is 0.35 [9]. The numerical analysis generates displacement and speed values from Eqs. 7 to 13, with similar analyses for motion in phases CE and EF. The terminal velocity is also calculated to ascertain any effects it has on the velocity.

2.4 Phase C

The main factors that affect the board-ball contact phases are the change in bounce and grip. The vertical energy loss parameter (i.e. coefficient of vertical restitution) is only considered at the first two points of board-ball contact at C and E. There is an established standard for regulated table tennis boards to generate a fixed rebound from a particular height [10]. We use this standard to establish the coefficient of

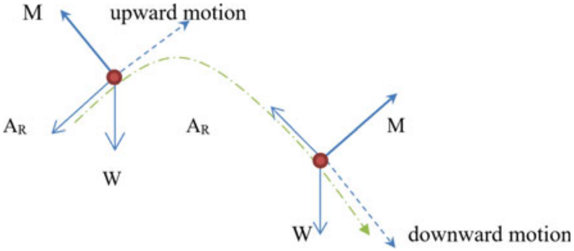


Fig. 2 Magnus lift of table tennis ball

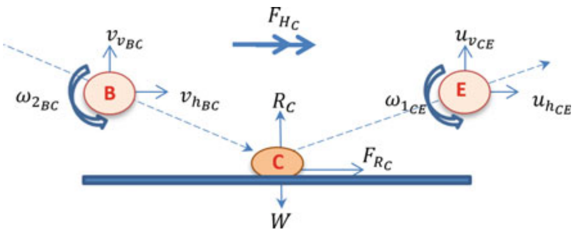


Fig. 3 Table tennis ball grips the surface at C

vertical restitution for our analysis. This parameter is represented by e_C and e_E at points C and E respectively, where $e_C = e_E$.

As the spinning ball grips the surface at C and E, changes in velocity and spin occur at these points [2]. This is illustrated in Fig. 3, where a ball of radius k is incident at an angle on a horizontal surface (board) with angular velocity $\omega_{2_{BC}}$, vertical velocity $v_{v_{BC}}$ and horizontal velocity $v_{h_{BC}}$. The ball rebounds off the surface with angular velocity $\omega_{1_{CE}}$, vertical velocity $u_{v_{CE}}$ and horizontal velocity $u_{h_{CE}}$.

The ball bends, grips and deforms upon impact with the surface, which affects the coefficient of horizontal restitution [2]. This coefficient is defined as the negated ratio of the horizontal speed of a point on the bottom of the ball with respect to the board after impact to the horizontal speed of a point on the bottom of the ball with respect to the board before impact. The coefficient of horizontal restitution has values ranging from -1 to 1, and can lead to the ball slowing down or even reversing direction. We assume that the net reaction force RC acts through the centre of the ball and thus did not affect the torque. It should however be pointed out that a torque is created when the frictional force F_{R_C} acts in a positive horizontal direction at the base of the ball. We also assume conservation of angular momentum about the point of contact between the ball and surface of the board. The ball is considered to be a thin hollow sphere when calculating its moment of inertia. Based on these assumptions for the rotating ball, Eqs. 17–20 are generated. A similar analysis is done for phase E, which is the other board-ball contact point.

2.5 Phase F

The displacement at point F at the end of the serve is instrumental in determining whether the serve is short, long, or qualified as a Ghost serve.

3 Results and Calculations

We consider vertical motion (↑) as positive and horizontal motion to the right (→) as positive.

3.1 Phase AB

The assumption of no losses in this phase allows the use of the Principle of Conservation of Energy. The final vertical velocity is

$$v_{v_{AB}} = -\sqrt{2gh_b} \tag{1}$$

where acceleration due to gravity is taken to be $g = 9.81\text{ms}^{-1}$. The toss was completely vertical, so the final horizontal velocity of the ball during this phase of motion is

$$v_{h_{AB}} = 0 \tag{2}$$

3.2 Phase B

We assume that the racket and ball coalesce, and that momentum is conserved, with the racket striking the ball at a tangent. The equations for the horizontal and vertical motion of the ball in this phase are

$$u_{h_{BC}} = \frac{m_r}{(m_r + m_b)} v_r \cos\theta \tag{3}$$

$$u_{v_{BC}} = \frac{m_r}{(m_r + m_b)} v_r \sin\theta + \frac{m_b}{(m_r + m_b)} v_{v_{AB}} \tag{4}$$

where mass of a raquet is taken as $m_r = 90$ g and the mass of the ball is given as $m_b = 2.7\,$g.

3.3 Analysis of the Airborne Phases

We assume that there is no rotational decrease during each airborne phase of motion, so that $\omega_i = \omega_f$. We also utilise $2 \leq v_r \leq 12\,\text{ms}^{-1}$, which is a realistic range of velocities for a legal serve with no spin. The corresponding range for Reynolds number is $5478 \leq \text{Re} \leq 32{,}868$. This is well within the Newton region which ranges from $103 \leq \text{Re} \leq 2.5 \times 10^5$, which in turn suggests that the drag coefficient in this region is approximately constant [1] at $C_D = 0.4$. It follows that air drag is

$$A_R = \frac{1}{2} C_D \rho_a \pi k^2 v^2 = q_d v^2 \tag{5}$$

The spinning ball (with angular velocity of 90 rads^{-1}) has a force acting on it due to the Magnus lift defined as

$$M = \frac{1}{2} C_L \rho_a \pi k^2 v^2 = q_m v^2 \tag{6}$$

since C_L is constant.

We assume that there is no rotational decrease during each airborne phase of motion, so that $\omega_i = \omega_f$. In our numerical approach, the initial values of horizontal distance $s_{h_0}$and velocity $\dot{s}_{h_0}$are known. Using Forward Euler and Central Differences to discretize the equations, we get

$$s_{h_1} = s_{h_0} + \dot{s}_{h_0} \Delta t \tag{7}$$

$$s_{h_{i+1}} = \ddot{s}_{h_i} \, (\Delta t)^2 + 2 s_{h_i} - s_{h_{i-1}} \tag{8}$$

$$\dot{s}_{h_i} = \frac{s_{h_{i+1}} - s_{h_i}}{\Delta t} \tag{9}$$

where $\ddot{s}_{h_i}$ is the discretized horizontal acceleration.

For the horizontal motion in Phase BC, the total horizontal force is given by

$$F_{h_{BC}} = -M_{h_{BC}} - A_{R_{h_{BC}}}$$

hence

$$m_b a_{h_{BC}} = -q_M v_{v_{BC}}^2 - q_D v_{h_{BC}}^2$$

$$\ddot{s}_{h_i} = \frac{-q_M \dot{s}_{v_i}^2 - q_D \dot{s}_{h_i}^2}{m_b} \tag{10}$$

and using (10) in (8)

$$s_{h_{i+1}} = \left(\frac{-q_M \dot{s}_{v_i}^2 - q_D \dot{s}_{h_i}^2}{m_b} \right) (\Delta t)^2 + 2 s_{h_i} - s_{h_{i-1}} \tag{11}$$

For the vertical motion in Phase BC

$$F_{v_{BC}} = M_{v_{BC}} - A_{R_{v_{BC}}} - W$$

hence

$$m_b a_{v_{BC}} = q_m v_{h_{BC}}^2 - q v_{v_{BC}}^2 - m_b g$$

and we obtain the discretized equations

$$\ddot{s}_{v_i} = \frac{q_m v_{h_{BC}}^2 - q v_{v_{BC}}^2 - m_b g}{m_b}$$

$$s_{v_{i+1}} = \left(\frac{-q_m\dot{s}_{h_i}^2 - q_d\dot{s}_{v_i}^2 - m_b g}{m_b}\right)(\Delta t)^2 + 2s_{v_i} - s_{v_{i-1}} \tag{12}$$

$$\dot{s}_{v_i} = \frac{s_{v_{i+1}} - s_{v_i}}{\Delta t} \tag{13}$$

Without loss of generality, the same procedures are repeated for the other two airborne phases with spin: Phases *CE* and *EF*.

3.4 Consideration of Terminal Velocity

Terminal velocity is calculated only with regards to the downward vertical motion from Newton's Second Law. Recall

$$F_{v_{BC}} = A_{R_{v_{BC}}} - W = q_d v_{v_{BC}}^2 - m_b g$$

Since $a_{v_{BC}} = 0$ at terminal velocity, it follows that $v_{t_{BC}}^2 = \frac{m_b g}{q_d} = 2.96\,\mathrm{ms}^{-1}$. From the numerical solution generated, the velocity of the Ma Lin ghost serve does not approach the terminal velocity. As such, this does not affect the solution.

3.5 Ball-Board Contact Analysis

Energy is lost vertically at the point of ball-board contacts at *C* and *E* [4]. The coefficient of vertical restitution [10] is $e_{v_C} = e_{v_E} = 0.77$. We also know that

$$\left|u_{v_{CE}}\right| = \left|e_{v_C} v_{v_{BC}}\right| \tag{14}$$

hence *C* and *E*

$$u_{v_{CE}} = -e_{v_C} v_{v_{BC}} \tag{15}$$

$$u_{v_{EF}} = -e_{v_B} v_{v_{CE}} \tag{16}$$

The ball deforms (slightly) and grips the surface at *C* and *E*, with the coefficient of horizontal restitution

$$e_{h_C} = -\frac{u_{h_{CE}-k\omega_{1CE}}}{v_{h_{BC}-k\omega_{i_{BC}}}}, \quad -1 < e_{h_C} < 1 \tag{17}$$

We assume that there is no rotational decrease during each airborne phase of motion, and $\omega_{i_{BC}} = 2\pi N_{BC}$. We also assume conservation of angular momentum about the point of contact with the ball and surface, hence

$$I\omega_1 + m_b k v_{h_{BC}} = I\omega_2 + m_b k u_{h_{CE}} \tag{18}$$

where $I = \alpha m_b k^2$ and $\alpha = \frac{2}{3}$ for a thin spherical sphere. We also assume that the frictional force $F_{R_C} = -m_b \frac{dv_h}{dt}$ and $F_{R_C} k = -I\frac{d\omega}{dt}$. Taking k to be constant in the bounce phase,

$$\int F dt = m_b \left(v_{h_{BC}} - u_{h_{CE}}\right) = m_b \alpha k \left(\omega_{1_{CE}} - \omega_{2_{BC}}\right)$$

These equations for spin are solved to obtain at C

$$\frac{u_{h_{CE}}}{v_{h_{BC}}} = \frac{1 - \alpha e_{h_C}}{1+\alpha} + \frac{\alpha\left(1 + e_{h_C}\right)}{1+\alpha}\frac{k\omega_{2_{BC}}}{v_{h_{BC}}} \tag{19}$$

$$\frac{\omega_{1_{CE}}}{\omega_{2_{BC}}} = \frac{\alpha - e_{h_C}}{1+\alpha} + \frac{1 + e_{h_C}}{1+\alpha}\frac{v_{h_{BC}}}{k\omega_{2_{BC}}} \tag{20}$$

Without loss of generality, there are similar equations generated at E. We consider the system for

$$e_{h_C} = e_{h_g} = 0.6$$

Thus, using the initial velocities and rotational speed at the beginning of each ball-board contact phase, the final values are generated.

4 Results

In the first airborne phase BC, and we incorporated the effect of air resistance and Magnus lift. The coefficients of drag and lift for these effects are held constant since the spin of the ball is assumed to be constant throughout this phase. The air resistance is proportional to the square of the velocity while the Magnus lift term is proportional to the square of the perpendicular velocity. Finite difference methods were utilized to generate numerical values for the ball displacement and velocity. This procedure was repeated for phases CE and EF.

The ball deforms (slightly) at both board-ball contact points at C and E. The resulting vertical and horizontal losses are incorporated into the analysis accordingly. The torque created upon impact of the ball with the board surface (as it grips the moving ball) converts some of the ball's translational energy into rotational energy. This significantly decreases the horizontal velocity of the ball, while simultaneously increasing the magnitude of the backspin at each of the board-ball contact points at C and E. These effects are cumulative, and may decrease the velocity of the ball to the point of reversing its trajectory at the second board-ball contact point E. In such cases, the Ma Lin short ghost serve is observed.

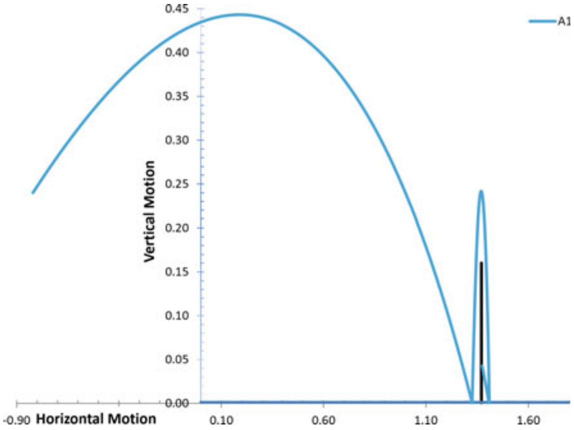


Fig. 4 Numerical solution of ball trajectory $A1$ depicting a Ma Lin ghost serve. Trajectory is measured in metres (m)

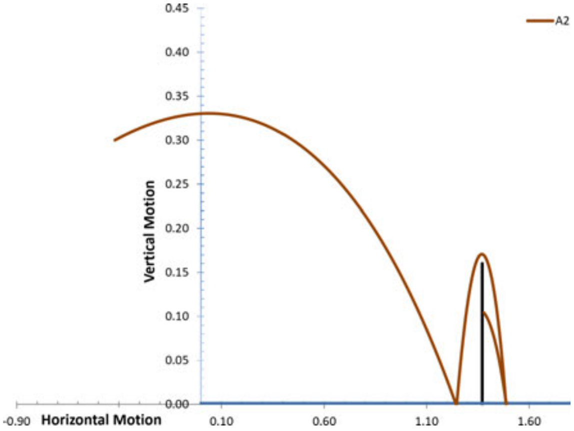


Fig. 5 Numerical solution of ball trajectory $A2$ depicting a Ma Lin ghost serve. Trajectory is measured in metres (m)

As players have individual playing techniques and styles of delivery, it is possible to generate Ma Lin ghost serves with widely ranging characteristics. We utilized our model to illustrate this by calculating the ball trajectories of four possible Ma Lin ghost serves by having one manipulated variable while keeping the backspin constant (as a controlled variable). The four resulting ball trajectories are illustrated in Figs. 4, 5, 6 and 7.

The Ma Lin ghost serve labelled $A1$ is representative of the original analysis done in the previous sections of this paper. Serve $A2$ was generated by reducing the serve angle (manipulated variable) previously adopted for serve $A1$. In order for the Ma Lin serve to be achieved with this variation, the height of toss, the height of contact and the distance of serve from the board were modified (responding variables). Serve $A3$ has a higher initial serve velocity (manipulated variable) than serve $A1$. In order to obtain the characteristics of a Ma Lin ghost serve for $A3$, alterations (responding variables) were made for the serve angle, height of racket contact and distance of serve from the board. Serve $A4$ varies from serve $A1$ by an increase in the height of the

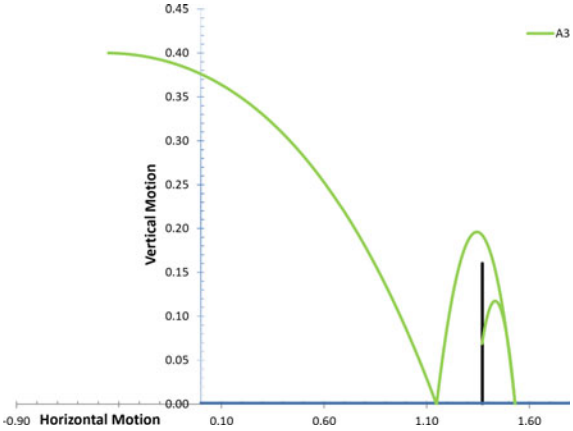


Fig. 6 Numerical solution of ball trajectory $A3$ depicting a Ma Lin ghost serve. Trajectory is measured in metres (m)

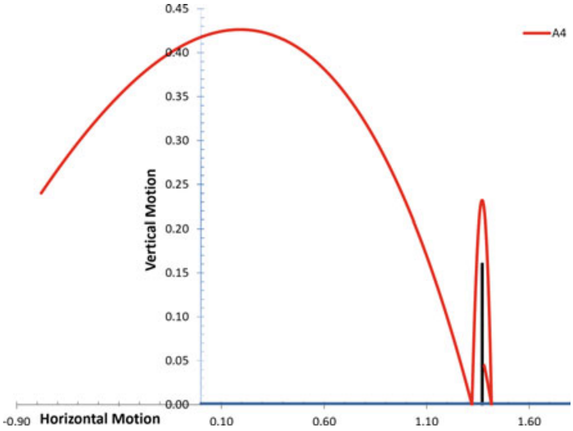


Fig. 7 Numerical solution of ball trajectory $A4$ depicting a Ma Lin ghost serve. Trajectory is measured in metres (m)

Table 2 Parameters used to generate the ball trajectories in Fig. 5

	Serve A_1	Serve A_2	Serve A_3	Serve A_4
v_r	5.5	5.5	6	5.5
θ	20	8	0	20
h_b	0.16	0.64	0.16	1.0
h_r	0.24	0.3	0.4	0.24
d	−0.82	−0.42	−0.45	−0.78

ball toss manipulated variable). The correction made to still generate a Ma Lin serve was an alteration in the serve position (responding variables). Table 2 summarizes the values of the parameters utilized to calculate the ball trajectories for serves $A1$ to $A4$.

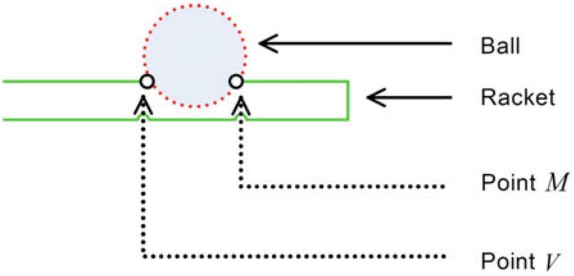


Fig. 8 Illustration of the indentation formed on the rubber racket surface upon contact with the ball

5 Discussion

Modern table tennis rackets are constructed with individual variations in adhesives, weight and surface textures which all have an impact on the ball trajectory. The layered hybrid rubber on the racket surface also preserves the table tennis ball from breaking upon impact [8]. For our purposes, we considered an ideal case where the principle of conservation of momentum could be used to determine the velocity after the racket-ball collision. We also assumed that the ball obtained a fixed rotational speed after impact with the surface of the racket.

Admittedly, several simplifying assumptions were made in the development of this model. The most significant of these assumptions are related to the dynamics of interaction between the racket and the ball. In postulating that the ball obtained a fixed rotational speed after impact with the surface of the racket, we assumed that the ball makes point contact with the racket at position B. In so doing, we avoided multiple complications that would arise from the surface of contact between the ball and racket.

Table tennis rackets are manufactured with a precise/delicate rubber covering, and as a result, each individual racket has a unique interaction with the ball. It is common knowledge among table tennis players that the softer the rubber on the racket, the easier it is to impart spin on the ball. When the ball is struck by the racket, an indentation is formed on the surface of the racket between the points M and V, as depicted in Fig. 8.

Spin is imparted on the ball during the deformation stage of the racket surface on impact, and that contact between the ball and the racket as the rubber reforms will reduce the energy transmitted to the ball. The contact reformation time is also dependent on the relative position/angle of the racket as it strikes the ball. We illustrate this with two examples when the ball is being served from the right end of the table (with the ball travelling to the left after service). In the first instance, as depicted in Fig. 9, clockwise backspin is imparted on the ball when the top of the racket is slanted towards the server with racket to the right of the ball, the ball will have a tendency to roll upwards along the racket surface when it is struck, which prolongs the time of surface contact between the ball and the racket.

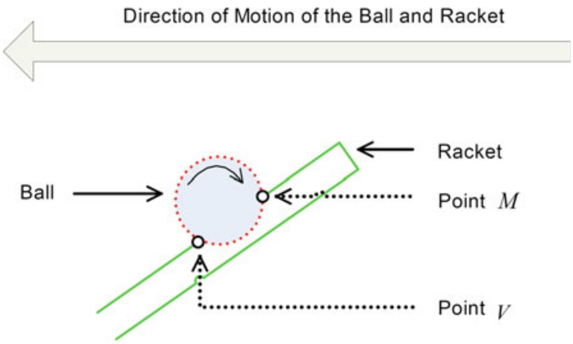


Fig. 9 Clockwise backspin imparted on the ball upon impact with the racket surface. In this configuration, the top of the racket is slanted toward the server and positioned to the right of the ball

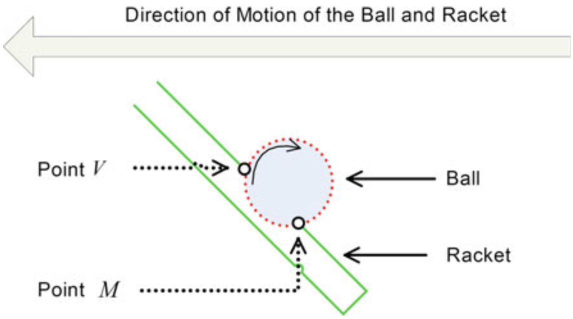


Fig. 10 Clockwise backspin imparted on the ball upon impact with the racket surface. In this configuration, the top of the racket is slanted away from the server and positioned to the left of the ball

In contrast, clockwise backspin can also be imparted with the top of the racket slanted away from the server and the racket to the left of the ball, the ball will have more of a tendency to slip off the racket, hence minimalising the time for surface contact, as shown Fig. 10.

Clearly, it would be advantageous to remove the simplifying assumption that the ball makes point contact with the surface of the racket at the time of the serve. The incorporation of a detailed surface contact analysis would therefore yield a much more accurate model for our purposes, and this will be the subject of our future investigations.

6 Conclusion

We successfully formulated a two-dimensional mathematical model for the Ma Lin ghost serve. Despite its inherent limitations, this paper represents a first attempt in the creation of a mathematical model for a relatively complicated backspin serve that is commonly employed by professional players in the sport of table tennis. The authors acknowledge that it would be greatly beneficial if experimental work was done in support of the model. This would allow us to obtain accurate values for all the required parameters. In the absence of this, we have estimated the necessary parameter values using established theoretical reference sources that have been duly cited. It should be noted however that the existence of a theoretical mathematical model is still of great value, as it represents an important first step in the development of a holistic model that describes the trajectory of the ball.

References

1. J. Cimbala, Drag on spheres, in *Lecture Notes 2012* (Penn State University, 2012), pp. 1–16. https://www.mne.psu.edu/cimbala/me325web_Spring_2012/Labs/Drag/intro.pdf
2. R. Cross, Grip-slip behaviour of a bouncing ball. Am. J. Phys. **70**, 1093–1102 (2002)
3. R. Cross, Backward bounce of a spinning ball. Eur. J. Phys. **39**, 045007 (2018)
4. A. Durey, R. Sevdel, Perfecting of a ball bounce and trajectories simulation software: in order to predict the consequences of changing table tennis rules. Int. J. Table Tennis **2**, 15–32 (1994)
5. M. Katsikadelis, T. Piliandis, N. Mantzourani, The interaction between serves and match wining table tennis players in the London 2012 Olympic Games, in *The 13th Sports Science Congress*, pp. 77–79 (2013)
6. H. Mohammed, R. Lakhan, D.M.G. Comissiong, A mathematical model for the Ma Lin table tennis ghost serve, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering*, 3–5 July, 2019 (London UK, 2019), pp. 36–41
7. K. Ou, P. Castonguay , A. Jameson, Computational sports aerodynamics of a moving sphere: simulating a ping pong ball in free flight, *29th AIAA Applied Aerodynamics Conference*, pp. 1–16 (2011)
8. R.G. Rinaldi, L. Manin, C. Bonnard, A. Drillon, H. Lourenco, N. Havard, Non linearity of the ball/rubber impact in table tennis: experiments and modeling. Procedia Eng. **147**, 348–353 (2016)
9. R. Sevdel, Determinant factors of the table tennis game—measurement and simulation of ball-flying curves. Int. J. Table Tennis Sci. **1**, 1–7 (1992)
10. *The International Table Tennis Handbook*, pp. 24–26 (2016)

Localization in Symplectic Geometry and Application to Path Integral and Supersymmetry

Philippe Durand

Abstract Equivariant geometry involves a group action on a manifold. This is the starting point to consider a super-geometry showing even and odd variables (bosons, fermions). The localization methods that provide source in the symplectic geometry (Duistermaat-Heckman formula), allow in certain cases to compute path integrals in an explicit way by using the concept of localization. Applications are important in topological field theory: They lead to the definition of new symplectics invariants.

Keywords Localization · Duistermaat-Heckman formula · Path integral · Quantum field theory · Supersymmetry · Gromov Witten invariants

1 Introduction

In this chapter,extended version of [1] presented in International Conference of Applied and Engineering Mathematics (ICAEM-2019) we discuss a fundamental problem, both in mathematics and physics. This is the problem of localization. Localization is a fundamental idea that makes a problem that is not countable a problem that it is, for example consider a map from a finite dimensional n vector space E to a finite dimensional m vector space F, the problem of finding the image of any vector of E by an application f in F is an uncountable problem because there is an infinite number of vectors in E thus an infinite number of possible images. Now if **we located** on the set of linear applications, to find the image of a vector of E, just know the image of n vectors: namely a basis of E, to know the image of any vector from the starting space. In other words, knowing a table of size $n \times m$: the matrix of the linear map is enough to solve the problem. We will begin by giving some examples of the problem of localization in mathematics, mainly in symplectic geometry, and quantum field theory, before then we recall the approach of the Feynman path

P. Durand (✉)
Département Mathématiques-Statistiques, Conservatoire National des Arts et Métiers,
292 rue Saint martin, 75141 Paris, France
e-mail: philippe.durand@lecnam.net

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_6

integral [2]. We then show how the introduction of fermionic variables can help to locate a problem and compute a path integral by using a localization principle. The main idea of the localization of the integrals comes from the oscillatory integrals, mainly the Laplace method on the localization around the critical points of a Morse function.

2 Motivations, Simple Examples: Quantization Problem of Feynman

The idea of localization has a lot of applications in quantum field theory (QFT). Feynmann has shown that the quantization of a classical field theory led to the computation of a path integral. This integral consists of calculating all the possible trajectories from a point A to a point B, then integrate on this space.

2.1 Toy Model

Let's start with a very simple problem: This problem is resolvable by a high school student: Consider a mouse moving over a grid consisting of n rows and m columns. His only possible actions are to advance or step up without ever going back or down. It is well known that counting all the possible trajectories consists of **counting the number of anagrams** consisting of n times the letter A and m times the letter M. In this case the sum of all possible trajectories is a finite number is the combinations of n among $n+m$: $N_T = \begin{pmatrix} n+m \\ n \end{pmatrix}$.

2.2 Path Integral

In the case of path integral, the set of trajectory is in general infinite. Localization problems then make perfect sense. Recall that the path integral introduced by Feynman is given by:

$$K(x(t_i), x(t_f)) = \int_{x(t_i)\to x(t_f)} \mathscr{D}x(t)exp\left(\frac{iS(x(t))}{\hbar}\right) \tag{1}$$

In this formula, the measure $\mathscr{D}$ is poorly defined: relates to an infinite number of paths from an initial configuration to a final configuration. The fact of considering forced passages by certain points leads to the notion of correlation function. Physicists have postulated that knowledge of all correlation functions makes it possible to understand a quantum field theory. In the case of our toy model, it is very easy to determine all the correlation functions passing through a point, two or more points of the grid).

3 Equivariant Cohomology, and Localization

Berline and Vergne [3] define equivariant cohomology. This leads to a very useful localization formula: The equivariant geometry localization formula. In the context of symplectic geometry, this formula becomes the formula of Duistermaat-Heckman. Witten was able to give a concrete application of this formula in the framework of supersymmetric field theory: We can then give a dictionary that allows us to translate the vocabulary of equivariant geometry into that of supersymmetric field theories.

3.1 Equivariant Differential Forms

Briefly, equivariant geometry consists in making a group G act on a variety. In the same way that the cohomology of de Rham is defined on M, equivariant cohomology can be defined on M equipped with G group action:

Denote $\mathbb{C}[\mathfrak{g}] \otimes \Omega^\bullet(M)$ the algebra of complex valued polynomial functions from $\mathfrak{g}$ to $\Omega^\bullet(M)$. The symbol $\mathfrak{g}$ denotes the Lie algebra associated with the G group (assumed to be a Lie group). An element here can be seen like a polynomial map α

$$\alpha : \mathfrak{g} \to \Omega^\bullet(M) : X \mapsto \alpha(X)$$

It will be denoted $\Omega^\bullet_G := (\mathbb{C}[\mathfrak{g}] \otimes \Omega^\bullet(M))_G$ the subalgebra of G, Invariant elements which are $\alpha \in \mathbb{C}[\mathfrak{g}] \otimes \Omega^\bullet(M)$ and satisfies:

$$\alpha(g \cdot X) = g \cdot \alpha(X), \forall g \in G, X \in \mathfrak{g}$$

An element in $\Omega^\bullet_G$ is called an equivariant differential form, the grading with value in $\mathbb{Z}$ on $\mathbb{C}[\mathfrak{g}] \otimes \Omega^\bullet(M)$ is defined by:

$$deg(P \otimes \alpha) = 2deg(P) + deg(\alpha), P \in \mathbb{C}[\mathfrak{g}], \alpha \in \Omega^\bullet(M)$$

3.2 Equivariant Cohomology

If it is necessary to define a differential $d_\mathfrak{g}$ in $\Omega^\bullet_G$ we set:

$$(d_\mathfrak{g}\alpha)(X) := d(\alpha(X)) - i_X\alpha(X) \tag{2}$$

where the contraction is given by, if $\omega \in \Omega^q(M)$ then

$$i_{X_M}\omega : \sum_j a_j t_j \mapsto i_{X_M}\omega := i_{\sum_j b_j \frac{\partial}{dx_j}}\omega = \sum_j t_j i_{b_j \frac{\partial}{dx_j}}\omega$$

And $d_{\mathfrak{g}}$ acts:

$$d_{\mathfrak{g}}\alpha)(X) = (P \otimes d\omega)(X) - (t_k P \otimes i_{b_k e_k})\omega(X)$$

We can see link between this differential and the cohomology operator BRST Q physicists.

3.3 Localization Formula in Equivariant Geometry

Let G a compact Lie group, with Lie algebra $\mathfrak{g}$, acting on an oriented compact manifold M of even dimension $2n$. Let α be an equivariantly closed form M. Let $X \in \mathfrak{g}$ and assume X_M has only isolated 0 then:

$$\int_M \alpha(X) = (-2\pi)^l \sum_{p \in M_0(X)} \frac{\alpha(X)_{[0]}(p)}{det^{\frac{1}{2}}(L_p)} \tag{3}$$

where $M_0(X)$ is the set of $p \in M$ with $X(p) = 0$, $\alpha(X)_{[0]}(p)$ design f function (0-form) at $p \in M$ and L_p is the transformation matrix of the action $\mathscr{L}_X \xi$ evaluated at $p \in M_0(X)$.

3.4 Localization Formula in Symplectic Geometry

A particularly interesting case of application of the above formula is that of symplectic geometry: given symplectic manifold (M, ω), hamiltonian function H, and hamiltonian vector field X_H, given G acting in (M, ω), and the moment map $\mu : M \to \mathfrak{g}^*$, we have The integral of Duistermaat-Heckman [4] theorem:

Theorem *Let M be a compact symplectic manifold of dimension $n = 2l$ and G a compact Lie group acting over it. Let $X \in \mathfrak{g}$ and let $X_M \in \mathfrak{X}(M)$ be the hamiltonian vector field generated by the moment map μ, that holds the identity $i_{X_M}\Omega = d\mu(X)$ with Ω be a symplectic 2-form over M. If $M_0(X)$ is the finite set of point on which X vanish, then*

$$\int_M e^{i\mu(X)} d\beta = i^l \sum_{p \in M_0(X)} \frac{(e^{i\mu(X)})_{[0](p)}}{det^{\frac{1}{2}}((L_p)(X))} \tag{4}$$

where the square root sign is chosen by canonical orientation on T_pM, we denote $d\beta := (e^{\frac{\Omega}{2\pi}})_{[n]} = \frac{\Omega^l}{(2\pi)^l l!}$, $n = 2l$.

Application 1
Let $S1 \circlearrowleft S2$ an action given by the rotation with respect to the vertical axis. Let us verify the hypothesis in the theorem of Duistermaat-Heckman for the action before:

1. $(S^2, sin\phi d\phi \wedge d\theta)$ is a symplectic manifold with $dim(S^2) = 2$.
2. S^1 is a compact Lie group.
3. The moment map μ is given by the height function, $\mu = cos\phi$.
4. X The vector field on the sphere is $X = \frac{\partial}{\partial\theta}{=}{-}y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y}$.
5. X The finite set of points where X vanishes, i.e., the fixed points of the action are $N = (0, 0, 1)$ and $S = (0, 0, -1)$.

We can apply the theorem of Duistermaat-Heckman!

1. The equivariant symplectic differential form is $\Omega_e q = \mu(X) + \omega_{S^2} = tcos\phi + sin\phi d\phi \wedge d\theta$.
2. The Lie algebra is given by $\mathfrak{g} = \mathbb{R} = t$. The vector field that this induces is $X_M = tX = t(-y\frac{\partial}{\partial x} + x\frac{\partial}{\partial y})$.
3. The transformation matrix A corresponding to the Lie bracket is given by $\mathscr{L}(X_M) : \mathfrak{X}(S^2) \to \mathfrak{X}(S^2), \xi \mapsto [X_M, \xi]$: $A = \begin{pmatrix} 0 & t \\ -t & 0 \end{pmatrix}$.
4. $det^{\frac{1}{2}}(A) = \pm t$, then the sign of the square root in the formula of Duistermaat-Heckman depends on the orientation coming from T_pS^1: At the point $(0, 0, 1)$ the sign is negative, and positive at $(0, 0, -1)$.
5. The Liouville form is $d\beta := \frac{\Omega^l}{(2\pi)^l l!} = \frac{1}{2\pi} sin\phi \wedge d\theta$.

By the theorem of Duistermaat-Heckman we obtain:

$$\int_{S^2} e^{itcos(\phi)} \frac{1}{2\pi} sin\phi d\phi \wedge d\theta = i\left(\frac{e^{itcos(0)}}{-t} + \frac{e^{itcos(\pi)}}{t}\right)$$

multiplying by 2π to both sides of the equality and expressing the exponential terms as a trigonometric function, it becomes to:

$$\int_{S^2} e^{itcos(\phi)} \frac{1}{2\pi} sin\phi d\phi \wedge d\theta = 4\pi \frac{sin(t)}{t} \tag{5}$$

the applications of the concept of localization goes beyond the framework of mathematics. We will now apply this concept to physics and particularly to the topological fields theories.

Application 2: Computation of the Volume of $\mathbb{CP}^k$
Let $S^1 \circlearrowleft S^{2k+1}$ defined by:

$$S^1 \circlearrowleft S^{2k+1} \to S^{2k+1} : e^{i\phi} \cdot (z_1, \ldots, z_{k+1}) \mapsto (e^{i\phi} \cdot z_1, \ldots, e^{i\phi} \cdot z_{k+1})$$

In this way the quotient $S^{2k+1}/S^1 \cong \mathbb{CP}^k$ can be express with the identification:

$$(z_1, \dots, z_{k+1}) \sim e^{i\phi} \cdot (z_1, \dots, z_{k+1})$$

Let $(e_i)_{i=1,k+1}$ be the canonical basis of $\mathbb{C}^{k+1}$. By means of the equivalence relation, the action can be written as:

$$(x_1, \dots, x_{k+1}, y_1, \dots, y_{k+1}) \in S^{2k+1} \subset \mathbb{R}^{2k+2} \mapsto [z_1 e_1 + \cdots z_{k+1} e_{k+1}] \in \mathbb{CP}^k \subset \mathbb{C}^{k+1}$$

Consider the symplectic form $\omega = \sum_{i=1}^{k+1} dx_i \wedge dy_i \in \mathbb{R}^{2k+2}$ restricted to S^{2k+1}, we have:

1. ω is invariant under the action of S^1: $(I_X \omega = 0)$
2. Ω: The pull back $q^*\Omega = \omega$ and identification of antipodal points, give a symplectic form: $\Omega = \sum_{i=1}^{k+1} dx_i \wedge dy_i$.

Consider now, the following action $g_{\overrightarrow{N}}$ that preserves the equivalence relation $S^{2k+1}/S^1 \simeq \mathbb{CP}^k$: $(\theta, (z_1, \dots, z_{k+1})) \in S^1 \times S^{2k+1} \mapsto (e^{in_1\theta} z_1, \dots, e^{in_{k+1}\theta} z_{k+1}) \in S^{2k+1}$, such that $\overrightarrow{N} = (n_1, \dots, n_{k+1}) \in \mathbb{N}^{2k+1}$. Then Choose n_1, n_{k+1} all different, if not the action is trivial. Now, take a vector field on S^{2k+1}:
Let $\overline{X} = \sum_{j=1}^{k+1} n_j(-y_j \frac{\partial}{\partial x_j} + x_j \frac{\partial}{\partial y_j})$ be a vector field of S^{2k+1}, and the map:

$$f : (z_1, \dots, z_{k+1}) \in S^{2k+1} \mapsto \frac{1}{2}(n_1|z_1|^2 + \cdots + n_{k+1}|z_{k+1}|^2)$$

1. f is well defined on $\mathbb{CP}^k$ by left invariance when is taken the quotient S^{2k+1}/S^1
2. $i_{\overline{X}}\omega = df$ (f is a moment map)
3. taken the pull back $\underline{X}$ of $\overline{X}$ on $\mathbb{CP}^k$, and the quotient of f denoted $\underline{f}$, then $i_{\overline{X}}\Omega = d\underline{f}$, by notation abuse, let $f = \underline{f}$.
4. Consider the equivariantly closed differential form on $\mathbb{CP}^k$: $\overline{\Omega}(t) := tf + \Omega$
5. we have $k+1$ fixed points on S^{2k+1} given by $e_j \pm = (0, \dots, 0, \pm 1, 0, \dots, 0)$
6. $f(e_j \pm) = \frac{1}{2} n_j$, $1 \leq j \leq k+1$

The fundamental vector field that induces S^1 over S^{2k+1} is:
$\overline{X} = t \sum_{j=1}^{k+1} n_j(-y_j \frac{\partial}{\partial x_j} + x_j \frac{\partial}{\partial y_j})$. Now we consider the Lie bracket:

$$\mathscr{L}(\overline{X}) : \xi \in \mathscr{X}(S^{2k}) \mapsto [\overline{X}, \xi] \in \mathscr{X}(S^{2k})$$

In the standard basis of the tangent of S^{2k} at a point p_i of the sphere, the matrix A of the transformation before by blocks:

$$A == \begin{pmatrix} \begin{pmatrix} 0 & -t(n_1 - n_j) \\ t(n_1 - n_j) & 0 \end{pmatrix} & & & \\ & \begin{pmatrix} 0 & -t(n_2 - n_j) \\ t(n_2 - n_j) & 0 \end{pmatrix} & & \\ & & \ddots & \\ & & & \begin{pmatrix} 0 & -t(n_{k+1} - n_j) \\ t(n_{k+1} - n_j) & 0 \end{pmatrix} \end{pmatrix}$$

Thus:

$$det[\mathscr{L}(\overline{X})_{p_j}]^{1/2} = det(A) = t^k \prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)$$

We can apply the localization formula to the integral:

$$\int_{\mathbb{CP}^k} e^{i\overline{\Omega}(t)} = \int_{\mathbb{CP}^k} e^{itf + i\Omega} = (-2\pi)^k \sum_{j=1}^{k+1} \frac{e^{\frac{1}{2}in_j t}}{t^k \prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)}$$

To calculate the volume of $\mathbb{CP}^k$, by expanding at both sides, for the left side:

$$\int_{\mathbb{CP}^k} e^{itf} (e^{i\Omega})_{[2k]} = \int_{\mathbb{CP}^k} e^{itf} \frac{(i\Omega)^k}{k!} = \int_{\mathbb{CP}^k} (1 + itf + \frac{(itf)^2}{2!} + \ldots) \frac{(i\Omega)^k}{k!}$$

for the right side:

$$\sum_{j=1}^{k+1} \frac{e^{\frac{1}{2}n_j t}}{t^k \prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)} = \sum_{j=1}^{k+1} \frac{1}{\prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)} \left(\frac{i}{t^k} + \cdots \frac{(\frac{1}{2}in_j)^k}{k!} + t \frac{(\frac{1}{2}in_j)^{k+1}}{(k+1)!} + \cdots \right)$$

The integral form of the volume form of $\mathbb{CP}^k$ is equal to the in the right hand side which is constant:

$$\int_{\mathbb{CP}^k} \frac{(\Omega)^k}{k!} = (-2\pi)^k \sum_{j=1}^{k+1} \frac{(\frac{1}{2}n_j)^k}{k! \prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)}$$

But, $\sum_{j=1}^{k+1} \frac{n_j^k}{\prod_{\substack{q=1 \\ q \neq j}}^{k+1} (n_q - n_j)} = 1$ thus:

$$\int_{\mathbb{CP}^k} \frac{(\Omega)^k}{k!} = \frac{\pi^k}{k!} \tag{6}$$

4 Become to Physic: Path Integral, Supersymmetry

4.1 Classical Fields

The concept of field is fundamental in physics. A field φ is a function of a world sheet (source space) into a target space, M (space physics) with a sufficient number of dimensions. So given a “package” (Σ, M, φ) and a classical action: S where: Σ is the source space, often a manifold: for The classical mechanic of the point is the time axis (world line), for the conformal field theories like strings theories: a Riemann surface ....

The Lagrangian density is a function on one or more fields and its first derivatives:

$$\mathscr{L} = \mathscr{L}(\varphi_1, \varphi_2, \ldots, \partial_\mu \varphi_1, \partial_\mu \varphi_2 \ldots) \tag{7}$$

The classical action is the integral of the classical Lagrangian density on the parameter space $S = \int \mathscr{L} d^{n+1}x$.

Principle of least action: The minimization of the action ($\delta S = 0$, leads to each field noted just φ to the Euler-Lagrange equation gives the equations of motion of the particle:

$$\frac{\partial \mathscr{L}}{\partial \varphi} - \partial_\mu \left(\frac{\partial \mathscr{L}}{\partial(\partial_\mu \varphi)} \right) = 0 \tag{8}$$

4.2 Example of Classical Fields

The free particle:

For a free particle, the field is simply the parameterized curve that describes the trajectory of the particle:$x(t)$ in free space. In this case, you can take to Lagrangian density

$\mathscr{L} = \mathscr{L}(x(t), \dot{x}(t)) = \frac{1}{2} m \dot{x}^2$, The Euler Lagrange equation reduce to:

$\frac{\partial \mathscr{L}}{\partial x} - \partial_t (\frac{\partial \mathscr{L}}{\partial(\dot{x})}) = -\partial_t (\frac{\partial \mathscr{L}}{\partial(\dot{x})}) = 0$ and the solution is uniform motion.

The free string:
For the free string, the field is simply the function that describes the trajectory of the string in the target space: $X(\tau, \sigma)$. Its Lagrangian density is contained in the Nambu-Goto action: $S = -T\int_{\tau_1}^{\tau_1} d\tau \int_0^l d\sigma \sqrt{\dot{X}^2 X'^2 - (\dot{X}X')^2}$ with $\dot{X}^2 = \dot{X}^\mu \dot{X}^\nu \eta_{\mu\nu}$, $\dot{X}X' = \dot{X}^\mu X'^\nu \eta_{\mu\nu}$
Euler Lagrange equation is then: $\partial_\tau \frac{\partial \mathscr{L}}{\partial \dot{X}^\mu} + \partial_\sigma \frac{\partial \mathscr{L}}{\partial X'^\mu} = 0$ and the solution is the equation of vibrating strings.

4.3 Noether Symmetries

Symmetry of the action:
The role of symmetry in physics is essential. We want, for example, such action invariant through a transformation like translation, rotation …: if $\varphi \rightarrow \varphi + \delta\varphi$ then $S \rightarrow S + \delta S$.
Noether's theorem:
Through any symmetry, the action is the same: $\delta S = 0$.
Example translation: $x \rightarrow x + \epsilon$ is taken up the free particle, ϵ small, independent of time,

$$\delta S = \int (m\dot{x}\dot{\epsilon})dt = \epsilon m\dot{x}|_{t_0}^{t_1} - \int (m\ddot{x})\epsilon dt$$

and as $\ddot{x} = 0$ we get: The symmetry by translation is equivalent to the conservation of momentum $p = m\dot{x}$.
Notation:
In physics, symmetry $x \rightarrow x + \epsilon$ is denoted $\delta x = \epsilon$.

4.4 Quantum Fields, QFT

Path integral:
Uncertainty on the position or momentum in quantum mechanic led to replace the classical solution (least action) by the partition function or the set of all possible solutions: It is the path integral:

$$\mathscr{Z} = \int_{\Sigma \rightarrow M} e^{-S(\varphi)} \mathscr{D}\varphi \tag{9}$$

Correlation Functions:
Similarly, one can calculate correlation functions, or functions with n points.

$$< \varphi_1(x_1), \ldots, \varphi_n(x_n) >= \int_{\Sigma \rightarrow M} \varphi_1(x_1) \ldots \varphi_n(x_n) e^{-S(\varphi)} \mathscr{D}\varphi \tag{10}$$

We can apply this machinery to the supersymmetric sigma model and define a new quantum field theory: the topological field theory *TFT* .The program developed by Witten is to calculate the correlation functions, by replacing each value by a cohomology class called *BRST* cohomology. That requires, introducing fermionic variables ***invariant*** under this ***generalized Noether symmetries***. These tools are ***supersymmetry***.

4.5 Supersymmetry

We can define a supersymmetric field theory $\Sigma \rightarrow M$ by adding fermionic variables, that is to say sections of some vector bundle E on Σ. A good image of a fermionic field is $\psi(x) = \Sigma f_i(x)dx_i$, a 1-form equipped with a wedge-product . We have the theorem:
Localization theorem: the path integral *is localized* around field configurations where fermionic variables stay invariant under supersymmetric transformations. Supersymmetric transformation is infinitesimal transformation of the action, which transforms bosons into fermions and vice versa.

4.5.1 Calculus Supersymmetric

We can define a supersymmetric Calculus:

Algebraic Computation: Let ψ_1, ψ_2 two fermions $\psi_1\psi_2 = -\psi_2\psi_1$ we deduce $\psi\psi = 0$ Let a bosonic variable *Xboson*, $\psi X = X\psi$
Calculus: $\int(a + b\psi)d\psi = b, \int \psi d\psi = 1, \int \psi_1\psi_2 \ldots \psi_n)d\psi_1 d\psi_2 \ldots d\psi_n = 1,$ $\int d\psi = 0$
Change of variables: We have: $\int \widetilde{\psi} d\widetilde{\psi} = \int \psi d\psi = 1.$

5 Localization in Physics

5.1 Example 1: Zero-Dimensional Supersymmetry

A “Toy” model is to make space for starting $\Sigma = \{P\}$ and target $M = \mathbb{R}$ the real line. In this context, a field is simply the variable x, the path integral is just $\mathscr{Z} = \int_M e^{-S(x)}dx$.

A supersymmetric action is given by:

$$S(x, \psi_1, \psi_2) = \frac{h'(x)^2}{2} - h''(x)\psi_1\psi_2.$$

hence the partition function:

$$\mathscr{Z} = \int e^{\frac{-h'(x)^2}{2} + h''(x)\psi_1, \psi_2} dx d\psi_1 d\psi_2$$

by developing in power series fermionic part, we get:

$$\mathscr{Z} = \int e^{\frac{-h'(x)^2}{2}} (1 + h''(x)\psi_1\psi_2) dx d\psi_1 d\psi_2,$$

but $\int d\psi = 0$, hence the first integral is zero, then:

$$\mathscr{Z} = \int\limits_M h''(x) e^{\frac{-h'(x)^2}{2}} dx \int \psi_1 d\psi_1 \int \psi_2 d\psi_2,$$

and as $\int \psi d\psi = 1$ (fermionic integration) we get:

$$\mathscr{Z} = \int\limits_M h''(x) e^{\frac{-h'(x)^2}{2}} dx$$

.

Supersymmetric Transformations

For the example above, we can define supersymmetric transformations that respect this action.

$$\delta x = \epsilon_1\psi_1 + \epsilon_2\psi_2$$
$$\delta\psi_1 = h'(x)\epsilon_2$$
$$\delta\psi_2 = -h'(x)\epsilon_1$$

We show that $\delta S = 0$, the fermionic variables are invariant for supersymmetric transformation iff $h'(x) = 0$. If $h'(x) \neq 0$, the change of variables $(x, \psi_1, \psi_2) \to (x - \frac{\psi_1\psi_2}{h'(x)}, \psi_1, \psi_2)$ shows that the partition function is zero outside the critical points. By expanding to second order near the critical point x_c

$$(h(x) = h(x_c) + \frac{h''(x_c)}{2}(x - x_c)^2) :$$
$$\mathscr{Z} = \int\limits_M h''(x) e^{\frac{-h'(x)^2}{2}} dx$$
$$\mathscr{Z} = \sum\nolimits_{h'(x_c)=0} h''(x_c) \int\limits_M exp\left(-\frac{(h''(x_c)(x-x_c))^2}{2}\right) dx,$$

with change of variables $y = |h''(x_c)|(x - x_c)$:

$$\mathscr{Z} = \sum_{h'(x_c)=0} \sqrt{\pi} \frac{h''(x_c)}{|h''(x_c)|}$$

Abstract

Supersymmetry: We just define an action for a supersymmetric field theory of dimension 0 by adding fermions, supersymmetry variables.
Invariance: This action is invariant under supersymmetric transformations.
Localization: The associated path integral is localized on the fields for which the fermions are invariant under supersymmetry.
Towards a generalization: This suggests defining an operator that vanishes on the fermionic fields. A fermionic field is associated to a differential form, there is an idea of cohomology below.

5.2 Exemple 2: Supersymmetric Quantum Mechanic (one Dimensional TQFT

1. Now we study a supersymmetric field theory in ***one dimension***. This is the model of ***supersymmetric quantum mechanics*** which has allowed Witten [5] to give a new proof of the ***index theorem*** [6, 7]. We considère the lagrangian:

$$\begin{aligned}&L = \tfrac{\dot{x}^2}{2} - \tfrac{h'(x)^2}{2} + i(\bar{\psi}\dot{\psi} - \dot{\bar{\psi}}\psi) - h''(x)\bar{\psi}\psi\\ &\psi = \psi_1 + i\psi_2\\ &\bar{\psi} = \psi_1 - i\psi_2\end{aligned}$$

2. Let $\pi = \frac{\partial L}{\partial \dot{\psi}} = i\bar{\psi}, p = \frac{\partial L}{\partial \dot{x}} = \dot{x}$ the conjugate moments
3. Let ***supersymmetric*** relations

$$\begin{aligned}&\delta_\epsilon x = \epsilon\bar{\psi} - \bar{\epsilon}\psi\\ &\delta_\epsilon \psi = \epsilon(i\dot{x} + h'(x))\ \epsilon = \epsilon_1 + i\epsilon_2\\ &\delta_\epsilon \bar{\psi} = \epsilon(i\dot{x} + h'(x))\end{aligned}$$

4. We can show:

$$\delta_\epsilon S = \int \delta L dt = \int \frac{d}{dt} L dt = 0.$$

5.2.1 Localization

The two operators of supersymmetry, are associated ***supercharge*** Q, $\bar{Q}$ with $Q^2 = \bar{Q}^2 = 0$ and we deduce an elliptic complex:

$$\mathscr{H}_F \xrightarrow{Q,\bar{Q}} \mathscr{H}_B \xrightarrow{Q,\bar{Q}} \mathscr{H}_F \xrightarrow{Q,\bar{Q}} \dots . \tag{11}$$

1. ***In hamiltonian*** formalism, $\{Q, \bar{Q}\} = 2H$
2. *SQM* compactified on S^1 give: $Tr(-1)^F e^{-\beta H} = dim\mathscr{H}^B_{(0)} - dim\mathscr{H}^F_{(0)}$ with F ***fermion number***.
3. The supertrace giving the index, expressed by: $Tr(-1)^F e^{-\beta H} = \int_{periodicBd} \mathscr{D}X \mathscr{D}\psi \mathscr{D}\overline{\psi} e^{-S}$
4. $\frac{\partial}{\partial\beta} Tr(-1)^F e^{-\beta H} = -\int_{periodicBd} \mathscr{D}X \mathscr{D}\psi \mathscr{D}\overline{\psi} He^{-S}$=0

5.2.2 Limite: From 1-Dim TQFT to 0-Dim TQFT

The fundamental result is that ***only time-independent contribute***: that reduce calculation to 0-dim TFT:

$$\mathscr{Z} = Tr(-1)^F e^{-\beta H} = \sum_{h'(x_c)=0} \sqrt{\pi} \frac{h''(x_c)}{|h''(x_c)|} \tag{12}$$

6 Example 3: A Model of Witten "A Side of the Mirror"

L, the supersymmetric lagrangian of a super-string is given by:

$$L = 2t \left(\int\limits_{\Sigma} \left(\frac{1}{2} g_{IJ} \partial_z \phi^I \partial_{\bar{z}} \phi^J \right) d^2z + \int\limits_{\Sigma} (i\psi_z^{\bar{i}} D_{\bar{z}} \chi^i g_{\bar{i}i} + i\psi_z^i D_z \chi^{\bar{i}} g_{\bar{i}i} - R_{i\bar{i}j\bar{j}} \psi_{\bar{z}}^i \psi_z^{\bar{i}} \chi^j \chi^{\bar{j}}) d^2z] \right) \tag{13}$$

The beginning of integral is the bosonic part of the action, the last , the fermionic part: fields are sections of bundles on Σ:

Fermionic part

- $\chi(z)$ a section $\mathscr{C}^\infty$ de $f^*TX \otimes \mathbb{C}$
- $\psi_z(z)$ a section $\mathscr{C}^\infty$ de $(T^{10}\Sigma)^* \otimes f^*T^{01}X$
- $\psi_{\bar{z}}$, a section $\mathscr{C}^\infty$ de $(T^{01}\Sigma)^* \otimes T^{10}X$

Supersymmetric transformation preserving action

$\delta x^I = \eta\chi^I \qquad \delta\chi^I = 0$
$\delta\psi_{\bar{z}}^i = \eta\partial_{\bar{z}}\phi_i \; \delta\psi_z^{\bar{i}} = \eta\partial_z\overline{\phi_i}$

- If $\delta\psi_{\bar{z}}^i = \delta\psi_z^{\bar{i}} = 0$, we recognize the conditions of Cauchy-Riemann!: The instantons of this model are curves "minimum energy" according to Gromov: holomorphic curves [8].

6.1 BRST Cohomology

At previous fermionic transformations one associates an operator Q (for charge), the terminology come from electromagnetism: charge is the integration of a "current". Mathematically, the operator Q has the properties of ordinary differential form (they will have an isomorphism between *BRST* cohomology with that of De Rham: we give now the main properties of this operator

Properties of the operator

- $Q(x^I) = \chi^I \; Q(\chi^I) = 0$
- Q is a linear operator.
- $Q(fg) = Q(f)g + fQ(g)$: Q is a derivation.

BRST cohomology

- We note that $Q^2 = 0$
- $H^p_{BRST} = \frac{KerQ:\mathscr{H}_p \to \mathscr{H}_{p+1}}{ImQ:\mathscr{H}_{p-1} \to \mathscr{H}_p}$ is the pth cohomology group BRST.

6.2 Correlation Functions BRST

In correlation functions fields are replaced by their cohomology classes [5], so they are defined modulo an exact term by:

Correlation Functions

Correlation functions of topological field theory will be given by:

$$< [\Phi_1(x_1)], \ldots, [\Phi_n(x_n)] >= \int\limits_{\Sigma \to M} \Phi_1(x_1) \ldots \Phi_n(x_n) e^{-S} \mathscr{D}x \mathscr{D}g \mathscr{D}\chi \mathscr{D}\psi \qquad (14)$$

Usually, the item $\Phi_i(x_i)]$ in topological field theorie is not a function but a cohomologie class which do not depend on the selected points on the Riemann surface.

Correlations functions of side A of the mirror

- Let $\omega_1, \ldots, \omega_n$ forms on M,

$$< [\omega_1], \ldots, [\omega_n] >= \int\limits_{\Sigma \to M} \omega_1 \ldots \omega_n e^{-(S_B(f)+S_F(f))} \mathscr{D}x \mathscr{D}g \mathscr{D}\chi \mathscr{D}\psi$$

- For the <u>theorem</u> location: the path integral is localized around holomorphic curves denoted $\tilde{f}$:

$$< [\omega_1], \ldots, [\omega_n] >= \int\limits_{\Sigma \xrightarrow{\tilde{f}} M} \omega_1 \ldots \omega_n e^{-(S_B(\tilde{f}))} \mathscr{D}x \mathscr{D}g \mathscr{D}\chi \mathscr{D}\psi$$

- $e^{-(S_B(\tilde{f}))} = e^{-\int_\Sigma \tilde{f}^* \omega}$ is a topological invariant gives the "degree" of application $\tilde{f}$.

6.3 Relationship with Enumerative Geometry

The path integral above can be rewritten:

$$< [\omega_1], \ldots, [\omega_n] >= \sum_{\beta \in H_2(M,\mathbb{Z})} e^{-\int_{\Sigma} \tilde{f}^{*}\omega} \int_{\tilde{f}(\Sigma)\in\beta} \omega_1 \ldots \omega_n \mathscr{D}x\mathscr{D}g\mathscr{D}\chi\mathscr{D}\psi \tag{15}$$

Here $\beta \in H_2(X,\mathbb{Z})$ is a cohomology class, “specifically” the degree of $\tilde{f}$.
Counting curves
We can hope the integral:

$$\int_{f(\Sigma)\in\beta} \omega_1 \ldots \omega_n \mathscr{D}x\mathscr{D}g\mathscr{D}\chi\mathscr{D}\psi$$

taken on a moduli space $\mathscr{M}$ to define properly, can provide an integer. This will be the case if the dimension of this moduli space is related to the number of fields $[\omega_i]$.
Gromov Witten invariants
These integrals, which give integers in good cases are Gromov Witten invariants [9–11]. Their knowledge provides a means of calculating correlation functions from a topological viewpoint and hope to understand better the physics!

7 Conclusion

Localization methods are crucial in mathematical physics. As we have seen, it allow t possible to make certain quantities calculable. It has brought back to the taste of the day some of the algebraic geometry methods as enumerative geometry. Its application to mathematical physics leads to the definition of moduli spaces and, in the best case, to instantons counting. This makes it possible to calculate certain correlation functions, resulting from a path integral. These methods have allowed a better understanding of quantum field theories in physics, they now connect topology, geometry and physic, to the concept of supersymmetry correctly defined from a mathematical point of view.

References

1. P. Durand, localization, path integral and supersymmetry, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering*, London, UK, 3–5 July 2019, pp. 1–5
2. Feynman, *Feynman’s Thesis (1942)—A New Approach to Quantum Theory, Laurie M. Brown* (Northwestern University, USA, 2005)

3. N. Berline, E. Getzler, M. Vergne, *Heat Kernels, and Dirac Operators* (Springer, Berlin Heidelberg, 1992)
4. J.J. Duistermaat, G.J. Heckman, On the variation in the cohomology of the symplectic form of the reduced phase space. Invent. Math. **69**(2), 259–268 (1982). https://doi.org/10.1007/BF01399506. MR 0674406
5. E. Witten, Topological quantum field theory. Commun. Math. Phys. **117**, 353–386 (1988)
6. M.F. Atiyah, I.M. Singer, The index of elliptic operators on compact manifolds. J. Bull. Amer. Math. Soc. **69**, 422–433 (1963)
7. H.B. Lawson, M.L. Michelsohn, *Spin Geometry* (Princeton University Press, Princeton, NJ, 1989)
8. D. McDuff, D. Salamon, *J-holomorphic Curves and Quantum Cohomology*. University Lecture Series, vol. 6 (A.M.S., 1994)
9. M. Kontsevich, Y. Manin, *Gromov-Witten Classes, Quantum Cohomology, and Enumerative Geometry*, Mirror Symmetry II, A.M.S/IP Studies in Advanced Mathematics, vol. 1 (A.M.S., 1997), pp. 607–653
10. M. Audin, *Cohomologie Quantique*, Séminaire Bourbaki, no. 807 (1995)
11. M. Audin, J.W. Morgan, P. Vogel, *Nouveaux invariants en géométrie et topologie*, Panoramas et Synthèses Numéro 11, 2001 S.M.F
12. D. Bennequin, *Monopôles de Seiberg-Witten et conjecture de Thom*, Séminaire Bourbaki, no. 807 (1995)
13. J.M. Bismuth, The Atiyah-Singer theorems: a probabilistic approach I. The index theorem. J. Funct. Anal., 56–99 (1984)
14. S. Donaldson, P. Kronheimer, *The Geometry of Four Manifolds* (Oxford University Press, Oxford, 1990)
15. T. Friedrich, *Dirac Operators in Riemannian Geometry*. Graduate Studies in Mathematics, vol. 25 (A.M.S., 2000)
16. P. Griffiths, J. Harris, *Principle of Algebraic Geometry* (Wiley, New York, 1978)
17. K. Hori, S. Katz, A. Klem, R. Pandharipande, R. Thomas, C. Vafa, R. Vakil, E. Zaslow, *Mirror Symmetry* (AMS, 2003)
18. D. McDuff, D. Salamon, *Introduction to symplectic géometry* (Oxford University Press, Oxford, 1995)
19. J. Le Potier, *fibrés vectoriels sur les courbes algébriques*, Publication de l'Université Paris 7
20. A.G. Sergeev, *Vortices and Seiberg Witten Equations*. Nagoya Mathematical Lecture, vol. 5 (2001)

A Simple Approach for Solving Linear Diophantine Equation in Two Variables

Yiu-Kwong Man

Abstract A simple approach for solving linear Diophantine equation in two variables is presented, which does not involve backward substitutions. Some illustrative examples are provided.

Keywords Elementary number theory · Euclidean algorithm · Extended euclidean algorithm · Linear diophantine equation · Non-backward substitution approach · Recurrence relations

1 Introduction

The Euclidean Algorithm and the Extended Euclidean Algorithm have important applications in areas such as number theory, discrete mathematics, computer algebra and matrix Pade approximations [1–4], etc. In this chapter, we introduce a simple approach for solving linear Diophantine equation in two variables, which does not involve backward substitutions as described in common textbooks of elementary number theory or discrete mathematics [2–5]. This approach is suitable for either hand calculation or computer programming.

The whole chapter is organized as follows. The common backward substitution approach is described in Sect. 2. Then, the new approach is introduced in Sect. 3, followed by some final remarks in the last section.

2 Backward Substitution Approach

Given two integers a and b, where $a > b > 0$. We can apply the Euclidean Algorithm to compute the greatest common divisor gcd(a, b) as follows:

Y.-K. Man (✉)
Department of Mathematics and Information Technology, The Education University of Hong Kong, 10 Lo Ping Road, New Territories, Hong Kong
e-mail: ykman@eduhk.hk

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_7

$$\begin{aligned} a &= bq_1 + r_1 \\ b &= r_1 q_2 + r_2 \\ r_1 &= r_2 q_3 + r_3 \\ &\vdots \\ r_{n-3} &= r_{n-2} q_{n-1} + r_{n-1} \\ r_{n-2} &= r_{n-1} q_n \end{aligned}$$

where q_{i}, r_{i} $(0 < i \leq n)$ denote the quotient and remainder of the *i*-th division. Then, $r_{n-1} = \gcd(a, b)$.

Suppose we want to find an integer solution of the linear Diophantine equation $ax + by = c$, where c is divisible by $\gcd(a, b)$. We can adopt a backward substitution approach, as illustrated in Example 1.

Example 1 Find an integer solution of the linear Diophantine equation 126x + 35y = 7.

Solution Applying the Euclidean Algorithm, we can obtain

$$\begin{aligned} 126 &= 35(3) + 21 \\ 35 &= 21(1) + 14 \\ 21 &= 14(1) + 7 \\ 14 &= 7(2) \end{aligned}$$

Hence, $\gcd(126, 35) = 7$. Applying backward substitutions to the above equations, we have

$$\begin{aligned} 7 &= 21 - 14(1) \\ &= 21 - [35 - 21(1)] \\ &= 21(2) - 35 \\ &= [126 - 35(3)](2) - 35 \\ &= 126(2) - 35(7) \end{aligned}$$

Thus, $x = 2$, $y = -7$ is an integer solution of 126x + 35y = 7.

From this example, we can see that the backward substitution approach could be quite tedious when there are more division steps involved in the Euclidean algorithm. In the next section, we will introduce an alternative approach for solving linear Diophantine equation in two variables, which can avoid backward substitutions.

3 Non-Backward Substitution Approach

Using the notations as above, we can express the successive remainders of the Euclidean algorithm as follows:

$$\begin{aligned} r_1 &= a - bq_1 \\ r_2 &= b - r_1q_2 = b - (a - bq_1)q_2 = -a\,q_2 + \ (1\ + q_1q_2)b \\ r_3 &= r_1 - r_2q_3 = \ (a - bq_1) - q_3[-a\,q_2 + \ (1\ + q_1q_2)b] \\ &= a(1\ + q_2q_3) - [q_1 + q_3(1\ + q_1q_2)]b \end{aligned}$$

$$\begin{aligned} r_4 &= r_2 - r_3q_4 \\ &= -aq_2 + \ (1\ + q_1q_2)b - q_4[a(1\ + q_2q_3) - [q_1 + q_3(1\ + q_1q_2)]b] \\ &= -a[q_2 + q_4(1\ + q_2q_3)] \ + \ [1\ + q_1q_2 + q_4[q_1 + q_3(1\ + q_1q_2)]]b \end{aligned}$$

and so on. According to the pattern of the coefficients of a and b shown in the above equations, we can find an integer solution of $ax + by = \gcd(a, b)$ by means of the following recurrence relations:

$$\begin{aligned} &x_0 = 0, x_1 = 1 \text{ and } x_i = x_{i-2} + q_ix_{i-1} \text{ for } 1 < i < n. \\ &y_0 = 1, y_1 = q_1 \text{ and } y_i = y_{i-2} + q_iy_{i-1} \text{ for } 1 < i < n. \end{aligned}$$

Let $x' = (-1)^n x_{n-1}$ and $y' = (-1)^{n-1} y_{n-1}$. Then, (x', y') is an integer solution of $ax + by = \gcd(a, b)$.

This approach has the advantage that the tedious task of backward substitutions can be avoided. Also, the recurrence relations concerned do not depend on the remainders r_i, but the quotients q_i only. It can enable us to compute an integer solution of the given Diophantine equation more conveniently.

Example 2 Find an integer solution of 126x + 35y = 7 by a non-backward substitution approach.

Solution Referring to Example 1, the successive quotients (except the last one) obtained are 3, 1 and 1, respectively. Using the recurrence relations described above, we have

i	0	1	2	3
q_i		3	1	1
y_i	1	3	4	7
x_i	0	1	1	2

Since $n = 4$, so $x = 2$, y $= -7$. In other words, (2, −7) is an integer solution of $126x + 35y = 7$, which is the same as that obtained in Example, but backward substitution is not required.

Example 3 Find the general solution of the Diophantine equation 2406x + 654y = 6 by a non-backward substitution approach.

Solution Applying the Euclidean algorithm, the successive quotients (except the last one) obtained are 3, 1, 2, 8 and 1, respectively. Using the recurrence relations described above, we have

i	0	1	2	3	4	5
q_i		3	1	2	8	1
y_i	1	3	4	11	92	103
x_i	0	1	1	3	25	28

Since $n = 6$, so $x = 28$, y $= -103$. In other words, (28, -103) is an integer solution of $2406x + 654y = 6$. Hence, the general solution is given by

$$x = 28 - 654t/6 = 28 - 109t \text{ and}$$
$$y = -103 + 2406t/6 = -103 + 401t,$$

where t is an arbitrary integer.

4 Final Remarks

A simple non-backward substitution approach for solving linear Diophantine equation in two variables has been introduced in this chapter. This work is a revised version of the paper presented at the International Conference of Applied and Engineering Mathematics (ICAEM 2019) held on 3–5 July 2019 at the Imperial College, London, UK [6]. Further works on exploring the possibility of extending similar ideas to handle linear Diophantine equation in n (>2) variables will be conducted in the near future. Since the Euclidean algorithm has wide applications in different areas, we anticipate that the non-backward substitution approach will be found useful for reference by teachers or researchers alike who are involved in teaching or conducting researches in elementary number theory, discrete mathematics or mathematical algorithms, etc.

References

1. P. Achuthan, S. Sundar, A new application of the Extended Euclidean algorithm for matrix Pade approximants. Comput. Math Appl. **16**, 287–296 (1988)
2. N. Biggs, *Discrete Mathematics*, Rev edn. (Oxford University Press, Oxford, 1990)
3. J. Gathen, J. Gerhard, *Modern Computer Algebra* (Cambridge University Press, Cambridge, 1999)

4. O. Ore, *Number Theory and Its History* (Dover, NY, 1988)
5. D.M. Burton, *Elementary Number Theory* (Allyn and Bacon, Boston, 1980)
6. Y.K. Man, A top-down approach for solving linear Diophantine equation, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2019*, 3–5 July 2019, London, UK, pp. 11–13

Transcomplex Integral

Tiago S. dos Reis and James A. D. W. Anderson

Abstract Transcomplex space is a compact metric space. The usual complex integral is defined in terms of complex numbers in Cartesian form but transcomplex numbers are defined in polar form and almost all transcomplex numbers, with infinite magnitude, have no Cartesian form. But eight infinite transcomplex numbers do have Cartesian form. We define the transcomplex integral as the limit of sums of these eight numbers. Thus we introduce the transcomplex integral.

Keywords Compactification · Transcomplex derivative · Transcomplex integral · Trans-complex number · Transmathematics · Transreal number

1 Introduction

Transcomplex space is a compact metric space [11]. The transcomplex circle at infinity, $\infty e^{i\theta}$, with real θ, compactifies both the complex and projective planes, and provides the extended-real compactification, $-\infty$ and ∞, of the real number line. Thus some of the usual compactifications of real, complex, and projective space are present in transcomplex space. This makes the transcomplex numbers inherently interesting—they already contain certain compactifications.

The Riemann Sphere is a unit sphere, conventionally arranged with its South Pole touching a plane. Infinite lines drawn through the North Pole, as the first point, and a second point on the sphere, are isomorphic with the complex plane. In the limit, lines passing through the North Pole and a circle of asymptotically zero radius, centred on the North Pole, describe the transcomplex circle at infinity and, by convention, a

T. S. dos Reis (✉)
Federal Institute of Education, Science and Technology of Rio de Janeiro,
Rio de Janeiro 27215-350, Brazil
e-mail: tiago.reis@ifrj.edu.br

J. A. D. W. Anderson
88 Lower Henley Road, Caversham, Reading RG4 5LE, England
e-mail: james.a.d.w.anderson@btinternet.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_8

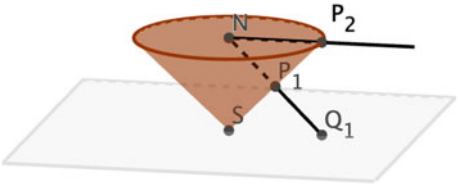


Fig. 1 Projective cone

line tangent to the North Pole is isomorphic to the point at nullity. Thus the Riemann Sphere approximates the transcomplex plane.

We obtain the transcomplex plane exactly, in Fig. 1, when we replace the sphere with a right cone with unit height and radius, whose apex, the South Pole, S, touches the plane and whose North Pole, N, is the centre of the base. A half-infinite line, projected from N, through a point, P_1, on the cone, south of the base, intersects the plane at Q_1. Thus the surface of the cone, south of the base, is isomorphic with the complex plane $re^{i\theta}$ with real r and θ. Projecting a half-infinite line from N, through a point P_2 on the base makes the base isomorphic with the circle at infinity, $\infty e^{i\theta}$ for real θ. A half-infinite line does not project from N to N. Thus choosing N as the second point produces an isolated point, N, that is isomorphic to the point at nullity, $\Phi e^{i\Phi}$, which, in turn, is isomorphic to the centre of projection [1]. Furthermore, if we identify points in the transcomplex plane with their reflections in the axis of the cone, we obtain the transprojective plane.

The transreal numbers [3, 13] totalise the real numbers by allowing division by zero in terms of three definite, non-finite numbers: negative infinity, $-\infty = -1/0$; positive infinity, $\infty = 1/0$; and nullity, $\Phi = 0/0$. In earlier work, real elementary functions and real limits were extended to transreal form [2, 8], as were both real differential and integral calculus [7, 9]. This extends real analysis to transreal analysis. Further to these works, a new transreal integral was developed [5].

We are now in the process of extending complex analysis to transcomplex analysis. Starting with the transcomplex numbers [6] the transcomplex topology, elementary functions and limits were developed [10, 11]. In the present paper we develop the transcomplex integral and just as much of the transcomplex derivative as we need. The current paper extends [12] by adding discussion of compactifications, providing new proofs, and two important corrections. This leaves a totalisation of the transcomplex derivative for future work, which will then extend complex analysis to transcomplex analysis. Thus the present paper can be seen as the penultimate step in extending complex analysis.

In order to understand this present paper, we advise the reader to review the transreal integral [5] and to review transcomplex numbers, their arithmetic, how their topology works, and how the elementary functions are defined on them [11].

The natural numbers have two different definitions, either including or excluding zero. The former definition is popular in Computer Science, the latter in Mathematics. Here we follow the mathematical convention $\mathbb{N} = \{1, 2, 3, \dots\}$.

2 Initial Considerations

In the complex domain, the integral along a curve is defined as follows. If $f : [a, b] \to \mathbb{C}$ is a function then, taking $u : [a, b] \to \mathbb{R}$ and $v : [a, b] \to \mathbb{R}$ such that $f = u + iv$, f is integrable in $[a, b]$ if and only if u and v are integrable in $[a, b]$ and the integral of f in $[a, b]$ is defined as $\int_a^b f(t)\,\mathrm{d}t := \int_a^b u(t)\,\mathrm{d}t + i \int_a^b v(t)\,\mathrm{d}t$. A smooth path is a differentiable function $\gamma : [a, b] \to \mathbb{C}$ such that γ' is continuous. Given a smooth path $\gamma : [a, b] \to \mathbb{C}$ and $f : \gamma([a, b]) \to \mathbb{C}$, f is integrable on γ if and only if $(f \circ \gamma)\gamma'$ is integrable in $[a, b]$ and the integral of f on γ is defined as $\int_\gamma f(z)\,\mathrm{d}z := \int_a^b f(\gamma(t))\gamma'(t)\,\mathrm{d}t$.

Notice that the definition of the complex integral is closely linked to the Cartesian form of complex numbers, $a + ib$ where $a, b \in \mathbb{R}$ and i is the imaginary unit. So we have a big problem to define the integral in the transcomplex domain. Since almost all infinite transcomplex number cannot be written as $a + ib$ with $a, b \in \mathbb{R}^T$, not all transcomplex functions can be represented by $u + iv$ with u and v being transreal functions.

Proposition 1 *For all $z \in \mathbb{C}^T_\infty$ if there are $a, b \in \mathbb{R}^T$ such that $z = a + bi$ then*

$$z \in \left\{ \frac{1}{0}, \frac{-1}{0}, \frac{i}{0}, \frac{-i}{0}, \frac{1+i}{0}, \frac{-1+i}{0}, \frac{-1-i}{0}, \frac{1-i}{0} \right\}.$$

Proof Let $z \in \mathbb{C}^T_\infty$ such that there are $a, b \in \mathbb{R}^T$ where $z = a + bi$.

Since $z \in \mathbb{C}^T_\infty$ there is $x \in \mathbb{C} \setminus \{0\}$ such that $z = \frac{x}{0}$ ([11], Remark 4).

Since $a, b \in \mathbb{R}^T$ there are $p, r \in \mathbb{R}$ and $q, s \in \{0, 1\}$ such that $a = \frac{p}{q}$ and $b = \frac{r}{s}$ ([13]) whence $a + bi = \frac{p}{q} + \frac{r}{s}\frac{i}{1}$. Since $\frac{r}{s}\frac{i}{1} = \frac{ri}{s}$ ([11], Definition 3 item d) we have that $z = a + bi = \frac{p}{q} + \frac{r}{s}\frac{i}{1} = \frac{p}{q} + \frac{ri}{s}$ that is

$$z = \frac{p}{q} + \frac{ri}{s}.$$

(i) The following do not occur simultaneously $p \neq 0$, $q \neq 0$, $r \neq 0$ and $s \neq 0$ because, otherwise, we would have $\frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs}$ ([11], Definition 3 item a) whence $z = \frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs}$. From $z = \frac{x}{0}$ and $z = \frac{ps+riq}{qs}$ we would have $\frac{x}{0} = \frac{ps+riq}{qs}$. Since $x, ps + riq \in \mathbb{C}$ and $0, qs \in \{0, 1\}$ there would be a positive $\alpha \in \mathbb{R}$ such that $0 = \alpha qs$ ([11], Definition 2) whence we would have $qs = 0$. Which is a contradiction.

(ii) If $p \neq 0$, $q \neq 0$, $r \neq 0$ and $s = 0$ then $\frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs}$ ([11], Definition 3 item a) whence $z = \frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs} = \frac{ri}{0}$. Since $r \in \mathbb{R} \setminus \{0\}$, denoting $\alpha = |r|$, we have that $\alpha \in \mathbb{R}$, α is positive and $ri = \alpha \frac{ri}{|r|}$ and $0 = \alpha \times 0$ whence $\frac{ri}{0} = \frac{\frac{ri}{|r|}}{0}$ ([11], Definition 2). Hence $z = \frac{ri}{0} = \frac{\frac{ri}{|r|}}{0} = \frac{\frac{r}{|r|}i}{0}$. Since $r \in \mathbb{R} \setminus \{0\}$ we have that either $\frac{r}{|r|} = -1$ or $\frac{r}{|r|} = 1$. Then either $\frac{\frac{r}{|r|}i}{0} = \frac{-i}{0}$ or $\frac{\frac{r}{|r|}i}{0} = \frac{i}{0}$. Thus either $z = \frac{-i}{0}$ or $z = \frac{i}{0}$.

(iii) The following do not occur simultaneously $p \neq 0$, $q \neq 0$, $r = 0$ and $s \neq 0$ for the same reason as item i.

(iv) The following do not occur simultaneously $p \neq 0$, $q \neq 0$, $r = 0$ and $s = 0$ because, otherwise, we would have $\frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs}$ ([11], Definition 3 item a) whence $z = \frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs} = \frac{0}{qs}$. From $z = \frac{x}{0}$ and $z = \frac{0}{qs}$ we would have $\frac{x}{0} = \frac{0}{qs}$. Since $x, 0 \in \mathbb{C}$ and $0, qs \in \{0, 1\}$ there would be a positive $\alpha \in \mathbb{R}$ such that $x = \alpha \times 0$ ([11], Definition 2) whence we would have $x = 0$. Which is a contradiction to the fact that $x \in \mathbb{C} \setminus \{0\}$.

(v) If $p \neq 0$, $q = 0$, $r \neq 0$ and $s \neq 0$ then $\frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs}$ ([11], Definition 3 item a) whence $z = \frac{p}{q} + \frac{ri}{s} = \frac{ps+riq}{qs} = \frac{p}{0}$. Since $p \in \mathbb{R} \setminus \{0\}$, denoting $\alpha = |p|$, we have that $\alpha \in \mathbb{R}$, α is positive and $p = \alpha \frac{p}{|p|}$ and $0 = \alpha \times 0$ whence $\frac{p}{0} = \frac{\frac{p}{|p|}}{0}$ ([11], Definition 2). Hence $z = \frac{p}{0} = \frac{\frac{p}{|p|}}{0}$. Since $p \in \mathbb{R} \setminus \{0\}$ we have that either $\frac{p}{|p|} = -1$ or $\frac{p}{|p|} = 1$. Then either $\frac{\frac{p}{|p|}}{0} = \frac{-1}{0}$ or $\frac{\frac{p}{|p|}}{0} = \frac{1}{0}$. Thus either $z = \frac{-1}{0}$ or $z = \frac{1}{0}$.

(vi) If $p \neq 0$, $q = 0$, $r \neq 0$ and $s = 0$ then $\frac{p}{q} + \frac{ri}{s} = \frac{\frac{p}{|p|} + \frac{ri}{|ri|}}{0}$ ([11], Definition 3 item a) whence $z = \frac{p}{q} + \frac{ri}{s} = \frac{\frac{p}{|p|} + \frac{ri}{|ri|}}{0} = \frac{\frac{p}{|p|} + \frac{r}{|r|}i}{0}$. Since $p \in \mathbb{R} \setminus \{0\}$ we have that either $\frac{p}{|p|} = -1$ or $\frac{p}{|p|} = 1$. Since $r \in \mathbb{R} \setminus \{0\}$ we have that either $\frac{r}{|r|} = -1$ or $\frac{r}{|r|} = 1$. Hence either $\frac{\frac{p}{|p|} + \frac{r}{|r|}i}{0} = \frac{-1-i}{0}$ or $\frac{\frac{p}{|p|} + \frac{r}{|r|}i}{0} = \frac{-1+i}{0}$ or $\frac{\frac{p}{|p|} + \frac{r}{|r|}i}{0} = \frac{1-i}{0}$ or $\frac{\frac{p}{|p|} + \frac{r}{|r|}i}{0} = \frac{1+i}{0}$. Thus either $z = \frac{-1-i}{0}$ or $z = \frac{-1+i}{0}$ or $z = \frac{1-i}{0}$ or $z = \frac{1+i}{0}$.

(vii) If $p \neq 0$, $q = 0$, $r = 0$ and $s \neq 0$ then in a similar way to the item v we have that either $z = \frac{-1}{0}$ or $z = \frac{1}{0}$.

(viii) The following do not occur simultaneously $p \neq 0$, $q = 0$, $r = 0$ and $s = 0$ for the same reason of the item iv.

(ix) The following do not occur simultaneously $p = 0$, $q \neq 0$, $r \neq 0$ and $s \neq 0$ for the same reason of the item i.

(x) If $p = 0$, $q \neq 0$, $r \neq 0$ and $s = 0$ then in a similar way to the item ii we have that either $z = \frac{-i}{0}$ or $z = \frac{i}{0}$.

(xi) The following do not occur simultaneously $p = 0$, $q \neq 0$, $r = 0$ and $s \neq 0$ for the same reason of the item i.

(xii) The following do not occur simultaneously $p = 0$, $q \neq 0$, $r = 0$ and $s = 0$ for the same reason of the item iv.

(xiii) The following do not occur simultaneously $p = 0$, $q = 0$, $r \neq 0$ and $s \neq 0$ for the same reason of the item iv.

(xiv) The following do not occur simultaneously $p = 0$, $q = 0$, $r \neq 0$ and $s = 0$ for the same reason of the item iv.

(xv) The following do not occur simultaneously $p = 0$, $q = 0$, $r = 0$ and $s \neq 0$ for the same reason of the item iv.

(xvi) The following do not occur simultaneously $p = 0$, $q = 0$, $r = 0$ and $s = 0$ for the same reason of the item iv.

□

From Proposition 1 we have it that only eight infinite transcomplex numbers can be written as $a + ib$ with $a, b \in \mathbb{R}^T$, namely, $\frac{1}{0}$, $\frac{-1}{0}$, $\frac{i}{0}$, $\frac{-i}{0}$, $\frac{1+i}{0}$, $\frac{-1+i}{0}$, $\frac{-1-i}{0}$ and $\frac{1-i}{0}$, which are: $\infty + i0$, $-\infty + i0$, $0 + i\infty$, $0 + i(-\infty)$, $\infty + i\infty$, $-\infty + i\infty$, $-\infty + i(-\infty)$ and $\infty + i(-\infty)$, respectively. Adding these eight numbers we can get an infinite number of infinite transcomplex numbers which, although they do not have Cartesian form, they are a sum of numbers which have Cartesian form. Note also these eight numbers are, in exponential form: ∞e^{i0}, $\infty e^{i\pi}$, $\infty e^{i\frac{\pi}{2}}$, $\infty e^{-i\frac{\pi}{2}}$, $\infty e^{i\frac{\pi}{4}}$, $\infty e^{i\frac{3\pi}{4}}$, $\infty e^{-i\frac{3\pi}{4}}$, $\infty e^{-i\frac{\pi}{4}}$, respectively. Summing numbers from these eight, we get numbers of the form $\infty e^{i\frac{l\pi}{2^n}}$ with $l, n \in \{0\} \cup \mathbb{N}$. Now notice that $\frac{l}{2^n}$, called dyadic rational numbers, are dense in $\mathbb{R}$. Therefore few infinite transcomplex numbers have Cartesian form but every infinite transcomplex number is the limit of a sequence of numbers which are sums of numbers which have Cartesian form.

The transcomplex integral, which we define here, is closely grounded in the above fact. For each transcomplex function f we take $(f_n)_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} f_n = f$ and f_n can be written as $\sum_{k=1}^{\infty} (u_k + iv_k)$ where u_k and v_k are transreal functions.

3 Integral on Transcomplex Numbers

A series of complex numbers is defined as the sequence $(s_n)_{n\in\mathbb{N}}$ where $s_n := \sum_{i=1}^{n} z_i = z_1 + \cdots + z_n$ and $(z_n)_{n\in\mathbb{N}} \subset \mathbb{C}$. We define transreal series in the same way [8]. But we need to be careful when defining series of transcomplex numbers because transcomplex addition is not associative. For example, $z_1 + z_2 + z_3$ is not well defined since $(z_1 + z_2) + z_3$ can be different from $z_1 + (z_2 + z_3)$ ([11], Proposition 21 item b).

Definition 1 Let $(z_n)_{n\in\mathbb{N}} \subset \mathbb{C}^T$. We define $\sum_{k=1}^{1} z_k := z_1$ and, for each $n \geq 2$, $\sum_{k=1}^{n} z_k := \left(\sum_{k=1}^{n-1} z_k\right) + z_n$. For each $n \in \mathbb{N}$ denote $s_n := \sum_{k=1}^{n} z_k$. The sequence $(s_n)_{n\in\mathbb{N}}$ is called a series of transcomplex numbers and is denoted by $\sum z_n$, each s_n is called a partial sum of $\sum z_n$ and z_n is called the nth term of $\sum z_n$. We say that $\sum z_n$ converges or is convergent if and only if there is the $\lim_{n\to\infty} s_n$. Otherwise, $\sum z_n$ diverges or is divergent. When $\sum z_n$ is convergent we denote $\sum_{k=1}^{\infty} z_k := \lim_{n\to\infty} \sum_{k=1}^{n} z_k$.

Definition 2 corrects Definition 2 of [12].

Definition 2 We denote

$$\mathcal{A} := \mathbb{C} \cup \{\Phi\} \cup \left\{\infty e^{i\frac{l\pi}{2^n}};\ l, n \in \{0\} \cup \mathbb{N}\right\}$$

and for each $z \in \mathcal{A}$ we define $\sum (a_k + b_k i)$, named *the Cartesian form of* z, in the following way:

(I) If $z \in \mathbb{C}$ then take $a, b \in \mathbb{R}$ such that $z = a + bi$ and define

$$\begin{aligned} a_1 &:= a \ \text{ and } b_1 := b \\ a_k &:= 0 \ \text{ and } b_k := 0 \ \text{ for all } k \geq 2 \end{aligned}$$

(II) If $z = \Phi$ then define

$$\begin{aligned} a_1 &:= \Phi \ \text{ and } b_1 := 0 \\ a_k &:= 0 \ \ \text{ and } b_k := 0 \ \text{ for all } k \geq 2 \end{aligned}$$

(III) If $z = \infty$ then define

$$\begin{aligned} a_1 &:= \infty \ \text{ and } b_1 := 0 \\ a_k &:= 0 \ \ \text{ and } b_k := 0 \ \text{ for all } k \geq 2 \end{aligned}$$

(IV) If $z \in \left\{\infty e^{i\frac{l\pi}{2^n}};\ l, n \in \{0\} \cup \mathbb{N}\right\} \setminus \{\infty\}$ then take $n \in \{0\} \cup \mathbb{N}$ and l odd with $l \in \{1, \ldots, 2^{n+1}\}$ such that $z = \infty e^{i\frac{l\pi}{2^n}}$ and for all $k \in \mathbb{N}$ define $a_k := a_k^{(n,\,l)}$ and $b_k := b_k^{(n,\,l)}$ where $\left(a_k^{(n,\,l)}\right)_{k\in\mathbb{N}}$ and $\left(b_k^{(n,\,l)}\right)_{k\in\mathbb{N}}$ are defined inductively in the following way:

For $n = 0$:

$$\begin{aligned} a_1^{(0,1)} &:= -\infty \ \text{ and } b_1^{(0,1)} := 0 \\ a_k^{(0,1)} &:= 0 \qquad \text{ and } b_k^{(0,1)} := 0 \ \text{ for all } k \geq 2 \end{aligned}$$

For $n = 1$:

$$\begin{aligned} a_1^{(1,1)} &:= 0 \ \text{ and } b_1^{(1,1)} := \infty \\ a_k^{(1,1)} &:= 0 \ \text{ and } b_k^{(1,1)} := 0 \quad \text{ for all } k \geq 2 \end{aligned}$$

and

$$\begin{aligned} a_1^{(1,3)} &:= 0 \ \text{ and } b_1^{(1,3)} := -\infty \\ a_k^{(1,3)} &:= 0 \ \text{ and } b_k^{(1,3)} := 0 \qquad \text{ for all } k \geq 2 \end{aligned}$$

For $n \geq 2$:

(i) if $0 \times 2^{n-2} < l \leq 1 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l)} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l)} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := \infty & \text{and } b_n^{(n,\,l)} := 0 & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(ii) if $1 \times 2^{n-2} < l \leq 2 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := 0 & \text{and } b_n^{(n,\,l)} := \infty & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(iii) if $2 \times 2^{n-2} < l \leq 3 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := 0 & \text{and } b_n^{(n,\,l)} := \infty & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(iv) if $3 \times 2^{n-2} < l \leq 4 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-2\times 2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-2\times 2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := -\infty & \text{and } b_n^{(n,\,l)} := 0 & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(v) if $4 \times 2^{n-2} < l \leq 5 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-2\times 2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-2\times 2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := -\infty & \text{and } b_n^{(n,\,l)} := 0 & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(vi) if $5 \times 2^{n-2} < l \leq 6 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-3\times 2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-3\times 2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := 0 & \text{and } b_n^{(n,\,l)} := -\infty & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(vii) if $6 \times 2^{n-2} < l \leq 7 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-3\times 2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-3\times 2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := 0 & \text{and } b_n^{(n,\,l)} := -\infty & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

(viii) if $7 \times 2^{n-2} < l \leq 8 \times 2^{n-2}$ then

$$\begin{array}{llll} a_k^{(n,\,l)} := a_k^{(n-1,\,l-4\times 2^{n-2})} & \text{and } b_k^{(n,\,l)} := b_k^{(n-1,\,l-4\times 2^{n-2})} & \text{for all } k \leq n-1 \\ a_n^{(n,\,l)} := \infty & \text{and } b_n^{(n,\,l)} := 0 & \\ a_k^{(n,\,l)} := 0 & \text{and } b_k^{(n,\,l)} := 0 & \text{for all } k \geq n+1 \end{array}$$

Remark 1 Notice that for any Cartesian form of a transcomplex number, say $\sum (a_k + b_k i)$, it follows that $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ are sequences of transreal numbers which have just a finite number of non-zero elements. Because of this, $\sum_{k=1}^{\infty}(a_k + b_k i)$ is nothing more than a finite sum. Hence we do not need to worry about convergence of series when we talk about the Cartesian form of a number.

Proposition 2 *Given $z \in \mathcal{A}$ and $\sum (a_k + b_k i)$ its Cartesian form, it follows that*

$$z = \sum_{k=1}^{\infty} (a_k + b_k i) \, .$$

Proof The result holds immediately from Definition 2 and from the properties of transcomplex arithmetic [11]. □

Definition 3 Let $D \subset \mathbb{C}^T$ and $f : D \to \mathbb{C}^T$ such that $f(D) \subset \mathcal{A}$. For each $w \in D$, denote the Cartesian form of $f(w)$ as $\sum (a_k(w) + ib_k(w))$. For each $k \in \mathbb{N}$, denote as u_k the function $u_k : D \to \mathbb{R}^T$ where $u_k(w) = a_k(w)$ for all $w \in D$ and as v_k the function $v_k : D \to \mathbb{R}^T$ where $v_k(w) = b_k(w)$ for all $w \in D$. Of course, $f = \sum_{k=1}^{\infty} (u_k + v_k i)$. We call $\sum (u_k + v_k i)$ the *Cartesian form of f*.

Remark 2 According to Remark 1 for any Cartesian form of a transcomplex number, say $\sum (a_k + b_k i)$, it follows that $\sum_{k=1}^{\infty} (a_k + b_k i)$ is actually a finite sum.

We correct [12] by noting that it is a different matter when we talk about the Cartesian form of a function. For example, let $f : [0, 1] \to \mathbb{C}^T$ where $f(t) = \infty e^{i\pi 2^{-\lfloor 1/t \rfloor}}$ for all $t \in [0, 1]$ (remember that $\lfloor x \rfloor$ denotes the floor function at the real number x and adopt the convention $\lfloor \infty \rfloor := \infty$). Denote the Cartesian form of the function f as $\sum (u_k + iv_k)$ and, for each $t \in [0, 1]$, denote the Cartesian form of the number $f(t)$ as $\sum (a_k(t) + ib_k(t))$.

Taking a fixed $t \in [0, 1]$, according to Definition 2, we have that $a_k(t) = 0$ and $b_k(t) = 0$ for all $k > \lfloor 1/t \rfloor$ whence $\sum_{k=1}^{\infty} (a_k(t) + ib_k(t)) = \sum_{k=1}^{\lfloor 1/t \rfloor} (a_k(t) + ib_k(t))$, that is, $\sum_{k=1}^{\infty} (a_k(t) + ib_k(t))$ is actually a finite sum—as it is stated in Remark 1.

However, according to Definition 2, $a_k\left(\frac{1}{k}\right)=\infty$ for all $k\in\mathbb{N}$ whence $u_k\left(\frac{1}{k}\right)=a_k\left(\frac{1}{k}\right)=\infty\neq 0$ for all $k\in\mathbb{N}$. Hence, for all $k\in\mathbb{N}$, there is $t\in[0,1]$ (for example, $t=\frac{1}{k}$) such that $u_k(t)\neq 0$ whence the function u_k is not identically zero. Thus $\sum_{k=1}^{\infty}(u_k+iv_k)$ is not a finite sum.

Although this difference matters, between the Cartesian form of a number and the Cartesian form of a function, we do not need to worry about convergence of series even when we talk about the Cartesian form of a function because we talk about pointwise convergence. In other words: $f=\sum_{k=1}^{\infty}(u_k+v_k i)$ is convergent if and only if $\sum_{k=1}^{\infty}(u_k(w)+iv_k(w))$ is convergent for each w in the domain of f. Since $\sum_{k=1}^{\infty}(u_k(w)+iv_k(w))$ is a finite sum for each w, it follows that $\sum_{k=1}^{\infty}(u_k(w)+iv_k(w))$ is convergent for each w whence $\sum_{k=1}^{\infty}(u_k+v_k i)$ is convergent.

Nonetheless, although we do not need to worry about convergence of $\sum(u_k+v_k i)$, we do need to worry about convergence of $\sum\left(\int_a^b u_k(t)\,\mathrm{d}t+i\int_a^b v_k(t)\,\mathrm{d}t\right)$. This is taken account of in Definition 6—as the reader will see.

Definition 4 For each $z\in\mathbb{C}^T$ we define $(z_n)_{n\in\mathbb{N}}$, named *the related sequence to* z, in the following way: If $z\in\mathcal{A}$ then define $z_n:=z$ for all $n\in\mathbb{N}$; if $z\notin\mathcal{A}$ then take $\theta\in(\pi,3\pi]$ such that $z=\infty e^{i\theta}$ and, for each $n\in\mathbb{N}$, take $l_n:=\max\left\{t\in\mathbb{N};\ \frac{t\pi}{2^n}<\theta\right\}$ and define $z_n:=\infty e^{i\frac{l_n\pi}{2^n}}$.

Proposition 3 *For all* $z\in\mathbb{C}^T$, *the related sequence to* z *converges to* z.

Proof Let $z\in\mathbb{C}^T$ be arbitrary. If $z\in\mathcal{A}$ then the result is immediate; if $z\notin\mathcal{A}$ then $z\in\mathbb{C}^T_\infty\setminus\left\{\infty e^{i\frac{l\pi}{2^n}};\ l,n\in\{0\}\cup\mathbb{N}\right\}$. Since $z\in\mathbb{C}^T_\infty$ there is $\theta\in(\pi,3\pi]$ such that $z=\infty e^{i\theta}$ ([11], first paragraph after the Proposition 14). Take $\left(\infty e^{i\frac{l_n\pi}{2^n}}\right)_{n\in\mathbb{N}}$, the related sequence to z. Notice that, by Definition 4, for all $n\in\mathbb{N}$, $\frac{l_n\pi}{2^n}<\theta\leq\frac{(l_n+1)\pi}{2^n}$ whence $0<\theta-\frac{l_n\pi}{2^n}\leq\frac{\pi}{2^n}$. Taking n tending to infinity in the latter inequality, we have that $\lim_{n\to\infty}\frac{l_n\pi}{2^n}=\theta$. Hence, since the function $g:\mathbb{R}\to\mathbb{C}$, where $g(x)=e^{ix}$ for all $x\in\mathbb{R}$, is continuous, we have that $\lim_{n\to\infty}e^{i\frac{l_n\pi}{2^n}}=e^{i\theta}$. Thus, by definition of the topology in $\mathbb{C}^T$ ([11], Proposition 22), we have that $\lim_{n\to\infty}\infty e^{i\frac{l_n\pi}{2^n}}=\infty e^{i\theta}$ whence $\lim_{n\to\infty}\infty e^{i\frac{l_n\pi}{2^n}}=\infty e^{i\theta}=z$. □

Definition 5 Let $D\subset\mathbb{C}^T$ and $f:D\to\mathbb{C}^T$ be arbitrary. We define $(f_n)_{n\in\mathbb{N}}$, named *the related sequence of functions to* f, in the following way: for each $w\in D$, take $(z_n)_{n\in\mathbb{N}}$, the related sequence to $f(w)$. For each $n\in\mathbb{N}$, define $f_n:D\to\mathbb{C}^T$ such that $f_n(w)=z_n$ for all $w\in D$.

Proposition 4 *Let* $D\subset\mathbb{C}^T$ *be arbitrary. For all* $f:D\to\mathbb{C}^T$, *the related sequence of functions to* f *converges uniformly to* f.

Proof Let $D\subset\mathbb{C}^T$, $f:D\to\mathbb{C}^T$ and $(f_n)_{n\in\mathbb{N}}$ be the related sequence of functions to f. Let positive $\varepsilon\in\mathbb{R}$ be arbitrary. For each $w\in D$ with $f(w)\notin\mathcal{A}$, denote $f(w)=\infty e^{i\theta(w)}$, where $\theta(w)\in(\pi,3\pi]$, and denote the related sequence to $f(w)$

as $\left(\infty e^{i\frac{l_n(w)\pi}{2^n}}\right)_{n\in\mathbb{N}}$. As the function $g : \mathbb{R} \to \mathbb{C}$, where $g(x) = e^{ix}$ for all $x \in \mathbb{R}$, is uniformly continuous in $[0, 3\pi]$, it follows that there is a positive $\delta \in \mathbb{R}$ such that $|e^{ix} - e^{iy}| < \varepsilon$ whenever $x, y \in [0, 3\pi]$ and $|x - y| < \delta$. Let $m \in \mathbb{N}$ such that $\frac{\pi}{2^m} < \delta$. It follows that if $n \geq m$ then $\left|\frac{l_n(w)\pi}{2^n} - \theta(w)\right| = \theta(w) - \frac{l_n(w)\pi}{2^n} \leq \frac{\pi}{2^n} \leq \frac{\pi}{2^m} < \delta$ for all $w \in D$ with $f(w) \notin \mathcal{A}$ whence $d(f_n(w), f(w)) = \left|\varphi(f_n(w)), \varphi(f(w))\right| = \left|\frac{1}{1+\frac{1}{\infty}} e^{i\frac{l_n(w)\pi}{2^n}} - \frac{1}{1+\frac{1}{\infty}} e^{i\theta(w)}\right| = \left|e^{i\frac{l_n(w)\pi}{2^n}} - e^{i\theta(w)}\right| < \varepsilon$ for all $w \in D$ with $f(w) \notin \mathcal{A}$ (the metric d and the homeomorphism φ are defined in [11], Proposition 22). Furthermore, $d(f_n(w), f(w)) = d(f(w), f(w)) = 0 < \varepsilon$ for all $n \in \mathbb{N}$ and for all $w \in D$ with $f(w) \in \mathcal{A}$. Whatever, if $n \geq m$ then $d(f_n(w), f(w)) < \varepsilon$ for all $w \in D$. □

Definition 6 Let $a, b \in \mathbb{R}$ with $a < b$ and $f : [a, b] \to \mathbb{C}^T$ such that $f([a, b]) \subset \mathcal{A}$ and take $\sum (u_k + iv_k)$ its Cartesian form. We say that f is *integrable in* $[a, b]$ if and only if u_k and v_k are integrable in $[a, b]$ for all $k \in \mathbb{N}$ and the series $\sum \left(\int_a^b u_k(t)\, \mathrm{d}t + i \int_a^b v_k(t)\, \mathrm{d}t\right)$ is convergent.[1] If f is integrable in $[a, b]$, the *integral of f in* $[a, b]$ is defined as

$$\int_a^b f(t)\, \mathrm{d}t = \sum_{k=1}^{\infty} \left(\int_a^b u_k(t)\, \mathrm{d}t + i \int_a^b v_k(t)\, \mathrm{d}t \right)$$

(the integrals $\int_a^b u_k(t)\, \mathrm{d}t$ and $\int_a^b v_k(t)\, \mathrm{d}t$ are defined in [5], Definition 3).

Definition 7 Let $a, b \in \mathbb{R}$ with $a < b$, $f : [a, b] \to \mathbb{C}^T$ and $(f_n)_{n\in\mathbb{N}}$ be the related sequence of functions to f. We say that f is *integrable in* $[a, b]$ if and only if f_n is integrable in $[a, b]$ for all $n \in \mathbb{N}$ and $\left(\int_a^b f_n(t)\, \mathrm{d}t\right)_{n\in\mathbb{N}}$ is convergent. If f is integrable in $[a, b]$, the *integral of f in* $[a, b]$ is defined as

$$\int_a^b f(t)\, \mathrm{d}t = \lim_{n\to\infty} \int_a^b f_n(t)\, \mathrm{d}t.$$

Remark 3 Notice that if f has Cartesian form then Definitions 6 and 7 give the same result.

Definition 8 Let $D \subset \mathbb{C}^T$. A *path in D* is a continuous function $\gamma : [a, b] \to D$ where $a, b \in \mathbb{R}$ and $a < b$. The image set of the function γ is denoted by $|\gamma|$.

Remark 4 For every path γ, either $|\gamma| = \{\Phi\}$ or $\Phi \notin |\gamma|$. Indeed, as γ is continuous, Φ is an isolated point ([11], first paragraph after the Proposition 27) and images of connected sets by continuous functions are connected ones, if $\Phi \in |\gamma|$ then $|\gamma| = \{\Phi\}$.

[1] We add the condition of convergence of the series $\sum \left(\int_a^b u_k(t)\, \mathrm{d}t + i \int_a^b v_k(t)\, \mathrm{d}t\right)$ to [12].

Now we define the derivative of a path. If $\gamma(t) \in \mathbb{C}$ then we have already the usual definition $\gamma'(t) = \lim_{h \to 0} \frac{\gamma(t+h)-\gamma(t)}{h}$. If $\gamma(t) = \Phi$ then $\gamma \equiv \Phi$ and we define $\gamma'(t) = \Phi$. We have a difficulty when $\gamma(t) \in \mathbb{C}_\infty^T$. If $\gamma(t) \in \mathbb{C}_\infty^T$ we have two possibilities: either there is a neighbourhood U of t such that $\gamma(U) \subset \mathbb{C}_\infty^T$ or $\gamma(U) \cap \mathbb{C} \neq \emptyset$ for all neighbourhoods U of t. In the first case, for all $s \in U$, $\gamma(s) = \infty e^{i\theta}$ for some $\theta \in \mathbb{R}$. Hence, as γ is continuous, $\gamma(U)$ is an arc of the circle at infinity. Thus there is a path β in $\mathbb{C}$ such that $\gamma(s) = \infty\beta(s)$ for all $s \in U$ and we define $\gamma'(t) = \infty\beta'(t)$. In the second case, $t \in \overline{\gamma^{-1}(\mathbb{C})}$ so if γ is differentiable in $\gamma^{-1}(\mathbb{C})$ we define $\gamma'(t) = \lim_{s \to t} \gamma'(s)$ if this limit exist.

Definition 9 Let $\gamma : [a, b] \to D \subset \mathbb{C}^T$ be a path and $t \in [a, b]$. Henceforth $\gamma'_{\mathbb{C}}(t)$ denotes the usual complex derivative of γ in t. We say that γ *is differentiable in t* if and only if one of the following conditions holds:

(i) $\gamma(t) \in \mathbb{C}$ and γ is differentiable in t in the usual sense. In this case we define the *derivative of* γ *in t* as the usual derivative of γ in t, that is, $\gamma'(t) := \gamma'_{\mathbb{C}}(t)$.
(ii) $\gamma(t) = \Phi$. In this case we define the *derivative of* γ *in t* as Φ, that is, $\gamma'(t) := \Phi$.
(iii) $\gamma(t) \in \mathbb{C}_\infty^T$ and there is a path β in $\mathbb{C}$ and a neighbourhood U of t such that $\gamma(s) = \infty\beta(s)$ for all $s \in U \cap [a, b]$. In this case we define the *derivative of* γ *in t* as $\infty\beta'(t)$, that is, $\gamma'(t) := \infty\beta'(t)$.
(iv) $\gamma(t) \in \mathbb{C}_\infty^T$ and $t \in \overline{E}$, where E is the set of all elements s from $[a, b]$ such that $\gamma(s) \in \mathbb{C}$ and γ is differentiable in s, and there is $\lim_{s \to t} \gamma'_{\mathbb{C}}(s)$. In this case we define the *derivative of* γ *in t* as $\lim_{s \to t} \gamma'_{\mathbb{C}}(s)$, that is, $\gamma'(t) := \lim_{s \to t} \gamma'_{\mathbb{C}}(s)$.

Definition 10 Let $\gamma : [a, b] \to D$ be a path. We say that γ is smooth when γ is differentiable and γ' is continuous in $[a, b]$.

Definition 11 Let $\gamma : [a, b] \to \mathbb{C}^T$ be a smooth path and $f : |\gamma| \to \mathbb{C}^T$ be a function. We say that f is *integrable on* γ if and only if $(f \circ \gamma)\gamma'$ is integrable in $[a, b]$. If f is integrable on γ, the *integral of* f *on* γ is defined as

$$\int_\gamma f(z)\, \mathrm{d}z := \int_a^b f(\gamma(t))\gamma'(t)\, \mathrm{d}t.$$

Since $\mathbb{C} \subset \mathbb{C}^T$, every complex function is a transcomplex function. Because of this the integral (in the transcomplex sense) is applicable to complex functions. In this way we now have two integrals for complex functions, namely the (transcomplex) integral defined in the present paper and the ordinary complex integral. However Proposition 5 shows that it is not a problem because the transcomplex integral and the ordinary complex integral are equivalent on complex functions.

Proposition 5 *Every complex function of complex variable is integrable in the usual sense if and only if it is integrable in the transcomplex sense. In other words: let* $\gamma : [a, b] \to \mathbb{C}$ *be a smooth path and* $f : |\gamma| \to \mathbb{C}$ *be a function, it follows that* f *is integrable on* γ *in the usual sense if and only if* f *is integrable on* γ *in the transcomplex sense. Furthermore both integrals have the same value.*

Proof Let $\gamma : [a, b] \to \mathbb{C}$ be a smooth path and $f : |\gamma| \to \mathbb{C}$ be a function. As $((f \circ \gamma)\gamma')([a, b]) \subset \mathbb{C}$, there are functions $u : [a, b] \to \mathbb{R}$ and $v : [a, b] \to \mathbb{R}$ such that $u + vi$ is the Cartesian form of $(f \circ \gamma)\gamma'$.

It follows that f is integrable on γ in the usual sense if and only if $(f \circ \gamma)\gamma'$ is integrable in $[a, b]$ in the usual sense if and only if u and v are integrable in $[a, b]$ in the usual sense if and only if u and v are integrable in $[a, b]$ in the transreal sense ([4], Theorem 2.2) if and only if, by Definition 6, $(f \circ \gamma)\gamma'$ is integrable in $[a, b]$ in the transcomplex sense if and only if, by Definition 11, f is integrable on γ in the transcomplex sense. And

$$\begin{aligned}\int\limits_{\substack{\gamma\\ \mathbb{C}}} f(z)\,\mathrm{d}z = \int\limits_{\substack{a\\ \mathbb{C}}}^{b} ((f \circ \gamma)\gamma')(t)\,\mathrm{d}t \qquad &= \int\limits_{\substack{a\\ \mathbb{R}}}^{b} u(t)\,\mathrm{d}t + i\int\limits_{\substack{a\\ \mathbb{R}}}^{b} v(t)\,\mathrm{d}t\\ = \int\limits_{a}^{b} u(t)\,\mathrm{d}t + i\int\limits_{a}^{b} v(t)\,\mathrm{d}t &= \int\limits_{a}^{b} ((f \circ \gamma)\gamma')(t)\,\mathrm{d}t\\ = \int\limits_{\gamma} f(z)\,\mathrm{d}z &\end{aligned}$$

where $\int_{\substack{\gamma\\ \mathbb{C}}} f(z)\,\mathrm{d}z$ denotes the integral of f on γ in the usual sense and $\int_{\substack{a\\ \mathbb{C}}}^{b}((f \circ \gamma)\gamma')(t)\,\mathrm{d}t$ denotes the integral of $(f \circ \gamma)\gamma'$ in $[a, b]$ in the usual sense and $\int_{\substack{a\\ \mathbb{R}}}^{b} u(t)\,\mathrm{d}t$ and $\int_{\substack{a\\ \mathbb{R}}}^{b} v(t)\,\mathrm{d}t$ denote, respectively, the integral of u and v in $[a, b]$ in the usual sense. □

Example 1 Let us calculate the integral of $z \mapsto |z|$ along the semi-straight line from 0 to ∞i, Fig. 2a.

Let $f : \mathbb{C}^T \to \mathbb{C}^T$ where $f(z) = |z|$ for all $z \in \mathbb{C}^T$ and $\gamma : [0, 1] \to \mathbb{C}^T$ where $\gamma(t) = \frac{t}{1-t}i$ for all $t \in [0, 1]$. Note that γ is continuous (it follows from the homeomorphism φ [11], Proposition 22) and continuously differentiable with $\gamma'(t) = \frac{1}{(1-t)^2}i$ for all $t \in [0, 1]$ (Definition 9). Thus $\displaystyle\int_{\gamma} f(z)\,\mathrm{d}z = \int_{0}^{1} f(\gamma(t))\gamma'(t)\,\mathrm{d}t =$

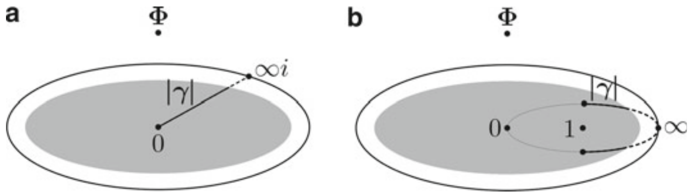


Fig. 2 **a** A semi-straight Line. **b** A semi-circle

$\int_0^1 \left|\frac{t}{1-t}i\right| \frac{1}{(1-t)^2}i \ \mathrm{d}t = \int_0^1 \frac{t}{1-t}\frac{1}{(1-t)^2}i \ \mathrm{d}t = i\int_0^1 \frac{t}{(1-t)^3} \ \mathrm{d}t = \infty i$ ([5], Definition 3).

Example 2 Let us calculate the integral of $z \mapsto \overline{z}$ along the circle at infinity. Let $f : \mathbb{C}^T \to \mathbb{C}^T$ where $f(z) = \overline{z}$ for all $z \in \mathbb{C}^T$ and $\gamma : [-\pi, \pi] \to \mathbb{C}^T$ where $\gamma(t) = \infty e^{it}$ for all $t \in [-\pi, \pi]$. Note that γ is continuous (it follows from the homeomorphism φ [11], Proposition 22) and continuously differentiable with $\gamma'(t) = \infty i e^{-it}$ (Definition 9). Thus $\int_\gamma f(z)\ \mathrm{d}z = \int_{-\pi}^{\pi} f(\gamma(t))\gamma'(t)\ \mathrm{d}t = \int_{-\pi}^{\pi} \overline{\infty e^{it}} \infty i e^{it}\ \mathrm{d}t = \int_{-\pi}^{\pi} \infty e^{-it} \infty i e^{it}$ $\mathrm{d}t = \int_{-\pi}^{\pi} \infty i e^{-it} e^{it} \quad \mathrm{d}t = \int_{-\pi}^{\pi} \infty i \quad \mathrm{d}t = i \int_{-\pi}^{\pi} \infty \quad \mathrm{d}t = \infty i \times 2\pi = \infty i$ ([5], Definition 3).

Example 3 Let us calculate the integral of $z \mapsto \frac{1}{|z|^2}$ along C, a semi-circle of centre 1 and radius $\frac{1}{2}$, Fig. 2b.

Notice that, for all $z \in C, d(z, 1) = \frac{1}{2}$ whence $|\varphi(z) - \varphi(1)| = \frac{1}{2}$, hence $\left|\varphi(z) - \frac{1}{2}\right| = \frac{1}{2}$ (the metric d and the homeomorphism φ are defined in [11], Proposition 22). As $d(z, 1) = \frac{1}{2}$ for all $z \in C$, we have that $\Phi \notin C$. Because of this, for all $z \in C$ there is $w \in \mathbb{C}$ such that $z = \varphi^{-1}(w)$. Thus C is made from points $\varphi^{-1}(w)$, with $w \in \mathbb{C}$, such that $\left|\varphi(\varphi^{-1}(w)) - \frac{1}{2}\right| = \frac{1}{2}$, that is, $\left|w - \frac{1}{2}\right| = \frac{1}{2}$. But $\left|w - \frac{1}{2}\right| = \frac{1}{2}$ if and only if $w = \frac{1}{2} + \frac{1}{2}e^{it}$ for some $t \in \mathbb{R}$. Therefore each point of C is given by $\varphi^{-1}\left(\frac{1}{2} + \frac{1}{2}e^{it}\right) = \frac{\left|\frac{1}{2}+\frac{1}{2}e^{it}\right|}{1-\left|\frac{1}{2}+\frac{1}{2}e^{it}\right|} e^{i\mathrm{Arg}\left(\frac{1}{2}+\frac{1}{2}e^{it}\right)} = \frac{1}{2-\sqrt{2+2\cos(t)}}(1 + \cos(t) + i\sin(t))$ for some $t \in \mathbb{R}$.

Let us take $\gamma : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to \mathbb{C}^T$, where $\gamma(t) = \frac{1}{2-\sqrt{2+2\cos(t)}}(1 + \cos(t) + i\sin(t))$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, and calculate the integral of $f : \mathbb{C}^T \setminus \{0\} \to \mathbb{C}^T$, where $f(z) = \frac{1}{|z|^2}$ for all $z \in \mathbb{C}^T \setminus \{0\}$, along $|\gamma|$. Firstly, note that γ is continuous. Indeed, if $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \setminus \{0\}$ then, clearly γ is continuous in t and, furthermore, $\gamma(0) = \infty$ ([11], Definition 3) and $\lim_{t\to 0}\gamma(t) = \infty$ (it follows from the homeomorphism φ [11], Proposition 22) whence γ is also continuous in 0. Secondly, note that γ is differentiable. In fact, clearly γ is differentiable in $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \setminus \{0\}$ with $\gamma'(t) = \frac{2\sin(t)\left(\sqrt{2+2\cos(t)}-4\right)}{4\left(2-\sqrt{2+2\cos(t)}\right)^2} + i\frac{8\cos(t)-\left(\sqrt{2+2\cos(t)}\right)^3}{4\left(2-\sqrt{2+2\cos(t)}\right)^2}$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \setminus \{0\}$ (Definition 9) and $\lim_{t\to 0}\gamma'(t) = \infty$ (it follows from the homeomorphism φ [11], Proposition 22) whence γ is differentiable in 0 with $\gamma'(0) = \infty$ (Definition 9). Furthermore γ' is continuous. Thus γ is a smooth path.

Now, notice that $f(\gamma(t)) = \frac{1}{|\gamma(t)|^2} = \frac{\left(2-\sqrt{2+2\cos(t)}\right)^2}{2+2\cos(t)}$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Thus $f(\gamma(t))\gamma'(t) = \frac{\sin(t)\left(\sqrt{2+2\cos(t)}-4\right)}{2(2+2\cos(t))} + i\frac{8\cos(t)-\left(\sqrt{2+2\cos(t)}\right)^3}{4(2+2\cos(t))}$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \setminus \{0\}$ and $f(\gamma(0))\gamma'(0) = \Phi$ ([11], Definition 3). Therefore

$$\begin{aligned}\int_\gamma f(z)\,\mathrm{d}z &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f(\gamma(t))\gamma'(t)\,\mathrm{d}t\\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\sin(t)\left(\sqrt{2+2\cos(t)}-4\right)}{2(2+2\cos(t))}\,\mathrm{d}t + i\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{8\cos(t)-\left(\sqrt{2+2\cos(t)}\right)^3}{4(2+2\cos(t))}\,\mathrm{d}t\\ &= (\pi-2-\sqrt{2})i\end{aligned}$$

([5], Definition 3).

Example 4 Let us calculate the integral of $z \mapsto z$ along a semi-circle at infinity. Let $f : \mathbb{C}^T \to \mathbb{C}^T$ where $f(z) = z$ for all $z \in \mathbb{C}^T$ and $\gamma : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \to \mathbb{C}^T$ where $\gamma(t) = \infty e^{it}$ for all $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. Note that γ is continuous and continuously differentiable with $\gamma'(t) = \infty i e^{-it}$ (Definition 9). Thus $\int_\gamma f(z)\,\mathrm{d}z = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f(\gamma(t))\gamma'(t)\,\mathrm{d}t = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \infty e^{it}\infty i e^{it}\,\mathrm{d}t = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \infty i e^{i2t}\,\mathrm{d}t = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \infty e^{i(2t+\pi)}\,\mathrm{d}t = \lim_{n\to\infty}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_n(t)\,\mathrm{d}t$ where, for each $t \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$, $(f_n(t))_{n\in\mathbb{N}}$ is the related sequence to $\infty e^{i(2t+\pi)}$.

Now, notice that, given $n \in \mathbb{N}$, from Definition 4 it follows that, $f_n(t) = \infty e^{i\frac{(l-1)\pi}{2^n}}$ for all $t \in \left[\frac{-2^n+l-1}{2^{n+1}}\pi, \frac{-2^n+l}{2^{n+1}}\pi\right)$, for each $l \in \{1, \dots, 2^{n+1}\}$. Hence, given $n \in \mathbb{N}$, denoting the Cartesian form of f_n as $\sum\left(u_k^{(n)} + iv_k^{(n)}\right)$ and denoting, for each $l \in \{1, \dots, 2^{n+1}\}$ and for each $t \in \left[\frac{-2^n+l-1}{2^{n+1}}\pi, \frac{-2^n+l}{2^{n+1}}\pi\right)$, the Cartesian form of $f_n(t) = \infty e^{i\frac{(l-1)\pi}{2^n}}$ as $\sum\left(a_k^{(n,\,l-1)} + ib_k^{(n,\,l-1)}\right)$, it follows that

$$\begin{aligned}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_n(t)\,\mathrm{d}t &= \sum_{k=1}^{\infty}\left(\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} u_k^{(n)}(t)\,\mathrm{d}t + i\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} v_k^{(n)}(t)\,\mathrm{d}t\right)\\ &= \sum_{k=1}^{\infty}\left(\sum_{l=1}^{2^{n+1}}\int_{\frac{-2^n+l-1}{2^{n+1}}\pi}^{\frac{-2^n+l}{2^{n+1}}\pi} a_k^{(n,\,l-1)}\mathrm{d}t + i\sum_{l=1}^{2^{n+1}}\int_{\frac{-2^n+l-1}{2^{n+1}}\pi}^{\frac{-2^n+l}{2^{n+1}}\pi} b_k^{(n,\,l-1)}\mathrm{d}t\right)\\ &= \sum_{k=1}^{\infty}\left(\sum_{l=1}^{2^{n+1}} a_k^{(n,\,l-1)} + i\sum_{l=1}^{2^{n+1}} b_k^{(n,\,l-1)}\right)\\ &= \sum_{k=1}^{\infty}(\Phi + i\Phi)\\ &= \Phi\end{aligned}$$

([5], Definition 3).

Therefore, $\int_\gamma f(z)\,\mathrm{d}z = \lim_{n\to\infty}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_n(t)\,\mathrm{d}t = \lim_{n\to\infty}\Phi = \Phi$.

Example 5 If $\gamma : [a, b] \to \mathbb{C}^T$ is the constant path $\gamma \equiv \Phi$ then $\int_\gamma f(z)\,\mathrm{d}z = \Phi$ for all$f : \{\Phi\} \to \mathbb{C}^T$. Indeed, $\int_\gamma f(z)\,\mathrm{d}z = \int_a^b f(\gamma(t))\gamma'(t)\,\mathrm{d}t = \int_a^b f(\Phi) \times \Phi\,\mathrm{d}t = \int_a^b \Phi\,\mathrm{d}t = \Phi$ ([5], Definition 3).

4 Conclusion

Earlier, real elementary functions, real limits and both real differential and integral calculus were extended to transreal forms. This extended real analysis to transreal analysis.

We now introduce the transcomplex integral and, incidentally, the derivative for transcomplex functions whose domain is a real interval. In future work, totalising the transcomplex derivative would complete the task of extending the main elements of complex analysis to transcomplex analysis.

Taking these results all together, a very large part of the computation of real and complex results can be done in the total systems of transreal and transcomplex numbers. This means that a very large part of practical computation can be done in an exception-free way, with beneficial consequences for safety-critical systems. In addition to their computational properties, the transcomplex numbers are also interesting because they form a compact metric space that contains some of the usual compactifications of the real number line and both the complex and protective planes.

References

1. J.A.D.W. Anderson, Representing geometrical knowledge. Philos. Trans. Roy. Soc. Lond. Ser. B **352**(1358), 1129–1139 (1997)
2. J.A.D.W. Anderson, T.S. dos Reis, Transreal limits expose category errors in IEEE 754 floating-point arithmetic and in mathematics, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014, WCECS 2014*, San Francisco, USA, 22–24 Oct 2014, pp. 86–91
3. J.A.D.W. Anderson, N. Völker, A.A. Adams, Perspex machine VIII: axioms of transreal arithmetic, in *Vision Geometry XV*, ed. by L. Jan Lateki, D.M. Mount, A.Y. Wu. Proceedings of SPIE, vol. 6499 (2007), pp. 2.1–2.12
4. T.S. dos Reis, Proper and improper Riemann integral in a single definition. Proc. Ser. Braz. Soc. Comput. Appl. Math. **5**, 010017-1–010017-7 (2016)
5. T.S. dos Reis, Transreal integral. Transmathematica 1–10 (2019). https://doi.org/10.36285/tm.v0i0.13
6. T.S. dos Reis, J.A.D.W. Anderson, Construction of the transcomplex numbers from the complex numbers, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2014, WCECS 2014*, San Francisco, USA, 22–24 Oct 2014, pp. 97–102

7. T.S. dos Reis, J.A.D.W. Anderson, Transdifferential and transintegral calculus, in *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering and Computer Science 2014, WCECS 2014*, San Francisco, USA, 22–24 Oct 2014, pp. 92–96
8. T.S. dos Reis, J.A.D.W. Anderson, Transreal limits and elementary functions. Transactions on Engineering Technologies—World Congress on Engineering and Computer Science (2014), pp. 209–225
9. T.S. dos Reis, J.A.D.W. Anderson, Transreal calculus. IAENG Int. J. Appl. Math. **45**(1), 51–63 (2015)
10. T.S. dos Reis, J.A.D.W. Anderson, Transcomplex topology and elementary functions, *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2016, WCE 2016*, London UK, 29 June–1 July 2016, pp. 164–169
11. T.S. dos Reis, J.A.D.W. Anderson, Transcomplex numbers: properties, topology and functions. Eng. Lett. **25**(1), 90–103 (2017)
12. T.S. dos Reis, J.A.D.W. Anderson, Integral on transcomplex numbers, *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2019*, London, UK, 3–5 July 2019, pp. 90–94
13. T.S. dos Reis, W. Gomide, J.A.D.W. Anderson, Construction of the transreal numbers and algebraic transfields. IAENG Int. J. Appl. Math. **46**(1), 11–23 (2016)

Arc Routing Based on the Zero-Suppressed Binary Decision Diagram

Renzo Roel P. Tan, Florian Sikora, Kazushi Ikeda, and Kyle Stephen S. See

Abstract A wide range of problems in operations research falls under arc routing problems, a domain which focuses on arc or edge features rather than node or vertex attributes. The undirected rural postman problem is a well-known problem in arc routing that seeks to determine a minimum cost walk that traverses a certain set of required edges on a given graph. The problem, arising in numerous real-world applications, is nondeterministic polynomial-time hard. The chapter presents a solution to the undirected rural postman problem based on the zero-suppressed binary decision diagram, a compact data structure that represents a family of sets. Through an extension to the frontier-based search method of diagram construction, the approach solves the problem by efficient enumeration, producing all feasible routes in addition to the optimal route. Instances of the problem put forward in literature are then solved as benchmark for the decision-diagram-based solution. As reasonable time is consumed, the method proves to be a practicable candidate in solving the undirected rural postman problem.

Keywords Arc routing · Enumeration algorithm · Frontier-based search · Graph optimization · Rural postman problem · Zero-suppressed binary decision diagram

R. R. P. Tan
Ateneo de Manila University, Quezon, Philippines

R. R. P. Tan (✉) · K. Ikeda
Nara Institute of Science and Technology, Nara, Japan
e-mail: rrtan@ateneo.edu; tan.renzo_roel_perez.tp7@is.naist.jp

K. Ikeda
e-mail: kazushi@is.naist.jp

F. Sikora
Paris Sciences et Lettres University, Paris, France
e-mail: florian.sikora@dauphine.psl.eu

K. S. S. See
Augmented Intelligence Pros, Incorporated, Manila, Philippines
e-mail: kyle1s@ai-pros.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_9

1 Introduction

Various problems in operations research may be posed as arc routing problems. See for example the following book [14]. Procedures such as mail delivery, garbage collection, and meter reading fall under the category. In general, arc routing consists of selecting an optimal route in a network, relying more on the edge properties than the vertex properties of the graph in the process.

A classic arc routing problem is the undirected rural postman problem [26], where the aim is to identify the route that covers a set of required edges with least cost—distance traveled, time consumed, *et cetera*. For instance, consider a postman who has to deliver the mail to a number of houses. The postman would have to traverse a subset of streets in the road network without missing any home that should receive mail. Such a circumstance may be formulated as an undirected rural postman problem where road intersections and road segments would be the vertices and edges of the graph, respectively. The required subgraph would then comprise sections of streets on which a house demands delivery.

Decision-diagram-based solutions have been proposed in combinatorial optimization over graphs. More specifically, basic routing problems on mass rapid transit systems were solved by a compressed representation called the zero-suppressed binary decision diagram [27]. In line with the above, the utility of the aforementioned variant of the binary decision diagram in resolving other discrete problems in graph-theoretic contexts is explored.

The research augments the frontier-based search in zero-suppressed binary decision diagram construction to accommodate the class-based constrained brought about by the problem having required and nonrequired edges. One then chooses instances of the undirected rural postman problem used across several sources for tests on computational efficiency. In closing, some advantages of the proposed zero-suppressed binary decision diagram method are established.

The chapter is organized as follows. The next three sections survey preliminary concepts on the undirected rural postman problem, the zero-suppressed binary decision diagram, and the frontier-based search. The fifth section details the methodology used in the study. The experiment results and the corresponding discussion are laid out in the sixth section. To conclude, a synthesis, as well as recommendations for future study, is contained in the last section.

For reference, the notation used throughout the paper is below.

V	The set of vertices $\{v_1, v_2, v_3, \ldots, v_{\lvert V\rvert}\}$.
E	The set of edges $\{e_1, e_2, e_3, \ldots, e_{\lvert E\rvert}\}$.
G	The graph (V, E) with vertex set V and edge set E.
R	The required set of edges $\{r_1, r_2, r_3, \ldots, r_{\lvert R\rvert}\}$.
U	A finite universe $\{x_1, x_2, x_3, \ldots, x_{\lvert U\rvert}\}$ with ordered elements.
$\mathcal{F}$	A family of subsets from U.
$\mathcal{D}$	The zero-suppressed binary decision diagram for $\mathcal{F}$.
$\mathbb{R}^+$	The set of nonnegative real numbers.
$f : A \to B$	A function mapping the set A to the set B.

To avoid confusion, one refers to the components of the given graph as vertices and edges and refers to the components of the diagram as nodes and arcs.

2 Undirected Rural Postman Problem

The notion of a walk in graph theory is requisite for the formulation of the problem.

Definition 1 (*Walk*) Given a graph, a walk is a sequence of edges $y_1, y_2, y_3, \ldots, y_p$ where y_k is the pair of vertices $\{w_k, w_{k+1}\}$ for $i = 1, 2, \ldots, p$.

In a walk, therefore, vertices and edges may be repeated in the sequence.

The undirected rural postman problem is first introduced in [26]. The problem, defined on a graph with required edges, involves designing a walk of least cost that traverses each edge in the required subgraph at least once. In contrast to the undirected Chinese postman problem [8] where the required subgraph is connected, the undirected rural postman would have no such guarantee [12]. The problem is proven to be nondeterministic polynomial-time hard [21]. The paper hires an uncomplicated definition of the problem suitable for the method of solution. The definition is motivated by [9].

Problem 1 (*Undirected Rural Postman Problem*) Given a graph $G = (V, E)$, a set of required edges $R \subset E$, a set of costs $C \subset \mathbb{R}^+$, and a cost function $c : E \to C$, find walk E' such that $\sum_{e \in E'} c(e)$ is minimum and $R \subseteq E'$.

A few applications from [9] are below.

- Street sweeping is a common example [2]. In a city, certain streets may require service more often than others.
- Snow plowing is also an application [13]. There are varying priority levels across categories of roads.
- Another example is school bus routing [1, 3, 7]. Students living in some street segments must board.

As there is prominence to the undirected rural postman problem, the different kinds of solutions have been frequently studied. Exact algorithms and heuristics exist in solving the undirected rural postman problem. Among the exact methods are a fundamental branch-and-bound scheme in [4], a cutting plane technique in [5], and a solution through an improved formulation in [10]. Heuristics that yield approximations are also used. The constructive approach by Frederickson [11] is most common; nevertheless, there are alternatives such as the Monte Carlo routine [6] and an improvement heuristics [15]. Recently, a solution based on genetic algorithms was also investigated [24].

3 Zero-Suppressed Binary Decision Diagram

The zero-suppressed binary decision diagram is a compact data structure capable of storing and manipulating families of sets [28]. Below is a formal definition of the zero-suppressed binary decision diagram [23].

Definition 2 (*Zero-Suppressed Binary Decision Diagram*) Consider a universe U. For $x_k \in U$, $x_i < x_j$ if and only if $i < j$. A zero-suppressed binary decision diagram is a labeled directed acyclic graph with the following properties.

1. Solely the root node has indegree 0.
2. Only two nodes with outdegree 0 called the 0-terminal and 1-terminal exist.
3. Each nonterminal node has exactly two outgoing arcs labeled by 0 and 1 and these are called the 0-arc and 1-arc, respectively.
4. The destination node of the 0-arc and the 1-arc of a node n is called the 0-child and 1-child of n, respectively.
5. Each nonterminal node is labeled by an element of U.
6. The label of a nonterminal node is strictly smaller than those of its children.

Clearly, a family of subsets $\mathcal{F}$ from U may be represented by a zero-suppressed binary decision diagram $\mathcal{D}$. For a subset $U' \in \mathcal{F}$, there is a path P from the root node to the 1-terminal of $\mathcal{D}$ to which U' corresponds [22]. A node labeled with x whose 1-arc is in P exists for all $x \in U'$. A theorem touching on the recursive structure of the diagram follows [23].

Theorem 1 *Let $\mathcal{D}$ correspond to $\mathcal{F}$ with root node e. The root node is either terminal or nonterminal.*

1. *If e is the 0-terminal then $\mathcal{F} = \emptyset$, the empty family.*
2. *If e is the 1-terminal then $\mathcal{F} = \{\emptyset\}$, the family containing only the empty set.*
3. *If e is nonterminal then it has two children. Let e_0 be the 0-child and e_1 be the 1-child of e. Denote the family with diagram rooted at e_i by $\mathcal{F}_i$. The family $\mathcal{F}$ may be written as $\mathcal{F}_0 \cup \left(\bigcup_{x \in \mathcal{F}_1} x \cup \{e\}\right)$.*

Connected to the 0-arc of e are the sets in $\mathcal{F}$ that do not contain e; connected to the 1-arc of e are the sets in $\mathcal{F}$ that do contain e. For emphasis, $\mathcal{F}_0 = \{x \mid x \in \mathcal{F}, e \notin x\}$ and $\mathcal{F}_1 = \{x \setminus \{e\} \mid x \in \mathcal{F}, e \in x\}$. Such an observation suggests the inherent recursiveness of the diagram structure. Operations on families of sets such as the intersection and union are conveniently done through the zero-suppressed binary decision diagram [23]. Pertinent results on the complexity of diagram operations are highlighted next [19].

Let $n(\mathcal{D})$ denote the number of nodes in diagram $\mathcal{D}$. For two diagrams $\mathcal{D}_1$ and $\mathcal{D}_2$ corresponding to families $\mathcal{F}_1$ and $\mathcal{F}_2$, computing for either $\mathcal{F}_1 \cap \mathcal{F}_2$ or $\mathcal{F}_1 \cup \mathcal{F}_2$ is of time complexity $\mathcal{O}(n(\mathcal{D}_1) \cdot n(\mathcal{D}_2))$.

The number of elements $|\mathcal{D}|$ in diagram $\mathcal{D}$ or equivalently, the number of subsets in $\mathcal{F}$ may be computed with time complexity $\mathcal{O}\left(n\left(\mathcal{D}\right)\right)$. Should items in U be weighted, the subset with maximum or minimum aggregate weight in $\mathcal{F}$ may be retrieved with the same cost. Lastly, the enumeration of the elements in $\mathcal{D}$ is of $\mathcal{O}\left(|\mathcal{D}| \cdot |U|\right)$ time complexity.

A unique reduced zero-suppressed binary decision diagram with the fewest nodes exists for a given family of sets [23]. An algorithm in [19] demonstrates reduction in linear time with respect to the number of nodes in the diagram.

Remark 1 A reduced zero-suppressed binary decision diagram is one that satisfies the properties below.

- There is no node whose 1-child is the 0-terminal.
- There are no distinct nodes that have the same label, 0-child, and 1-child.

In combinatorial optimization, feasible solutions stemming from certain set-ups are often subsets that may be stored in a zero-suppressed binary decision diagram [17]. Aside from extracting the optimal solution, the number of unique solutions, the mean and variance of solutions, and even an enumeration of solutions are easily obtained [19]. For a problem on graphs, a zero-suppressed binary decision diagram may view the edges of the graph as items in a universe. Each node in the diagram is labeled with an edge and indicates the inclusion of the edge. A path from the root node to the 1-terminal represents a subgraph and the diagram would then correspond to a collection of subgraphs.

The utility of the zero-suppressed binary decision diagram in familiar problems from combinatorics serve as guide for the interpretation of a diagram. The combination problem is the problem of finding ways in choosing k objects from n unique objects. The diagram generated for the case of $k = 3$ and $n = 6$ is explained in Fig. 1. Furthermore, the common knapsack problem is tackled in Fig. 2 through an example from [20]. The goal is to identify which selection of items with aggregate weight not exceeding the limit would have the maximum aggregate value.

4 Frontier-Based Search

A framework for the construction of the zero-suppressed binary decision diagram is the frontier-based search [18]. Diagrams representing sets of subgraphs that are paths, matchings, and trees, among others, may be generated depending on the problem setting. The section provides a summary of the method for representing a set of walks in a graph based on [27].

Take into consideration an undirected graph $G = (V, E)$ with weighted edges. It is given that G is simple and connected. An element of E defined by a 2-subset of V is a unique edge in G. A subgraph of G may thus be defined as $G' = \left(V', E'\right)$,

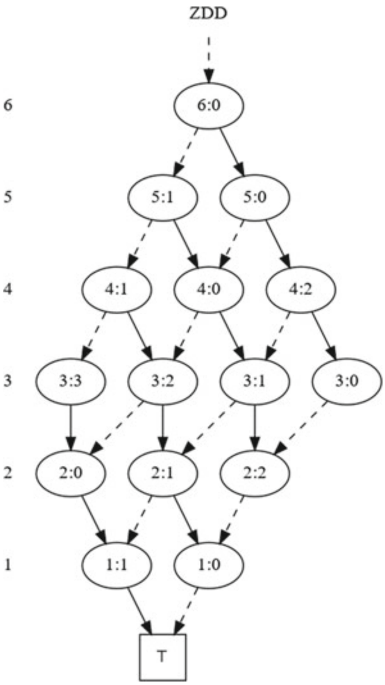


Fig. 1 The diagram for the combination problem of choosing 3 items from a set of 6 items: the paths from the root node ending in the 0-terminal are dropped. The 1-arcs and 0-arcs are represented by the solid and broken lines, respectively. A 3-subset of the cardinality 6 universe corresponds to a traversal from the root node to the 1-terminal. At each level, one proceeds to the 1-child if the item is present and to the 0-child if the item is not

where $V' = \bigcup_{e \in E'} e$ and $E' \subseteq E$. To clarify, a union $\bigcup_{i=1}^{j} e_i$ is the set of vertices to which at least one of $e_1, e_2, e_3, \ldots, e_j$ is incident. For the purpose of the study, there is no vertex with degree 0 in any subgraph. Let the universe for the zero-suppressed binary decision diagram be E, with ordered elements $e_1 < e_2 < e_3 < \cdots < e_{|E|}$.

The frontier-based search constructs the diagram breadth-first, starting with labeling the root node as e_1. Nodes labeled e_{i+1} are created only after all nodes labeled e_i have been produced for $i = 1, 2, 3, \ldots, |E| - 1$. Both the 1-arc and the 0-arc of a node must have the 1-terminal, the 0-terminal, or a node labeled e_{i+1} as destination. Simultaneously, subgraph specifications are stored through an array `n.deg` cached in every node n. The said array maps a particular subset of V to the set of natural numbers. Two nodes n and n' are merged if `n.deg` is equal to `n'.deg`. This strategy is aptly called node sharing [18].

Let k be the largest edge index for an edge incident to vertex v. One tags the node v as fixed whenever e_k is done being processed [18]. In other words, `deg[v]` is no longer updated because it is independent of $e_{k+1}, e_{k+2}, e_{k+3}, \ldots, e_{|E|}$. For $j = 1, 2, 3, \ldots, |E| - 1$, the jth frontier is defined in [18] as

$$F_j = \left(\bigcup_{i=1}^{j} e_i \right) \cap \left(\bigcup_{i=j+1}^{|E|} e_i \right)$$

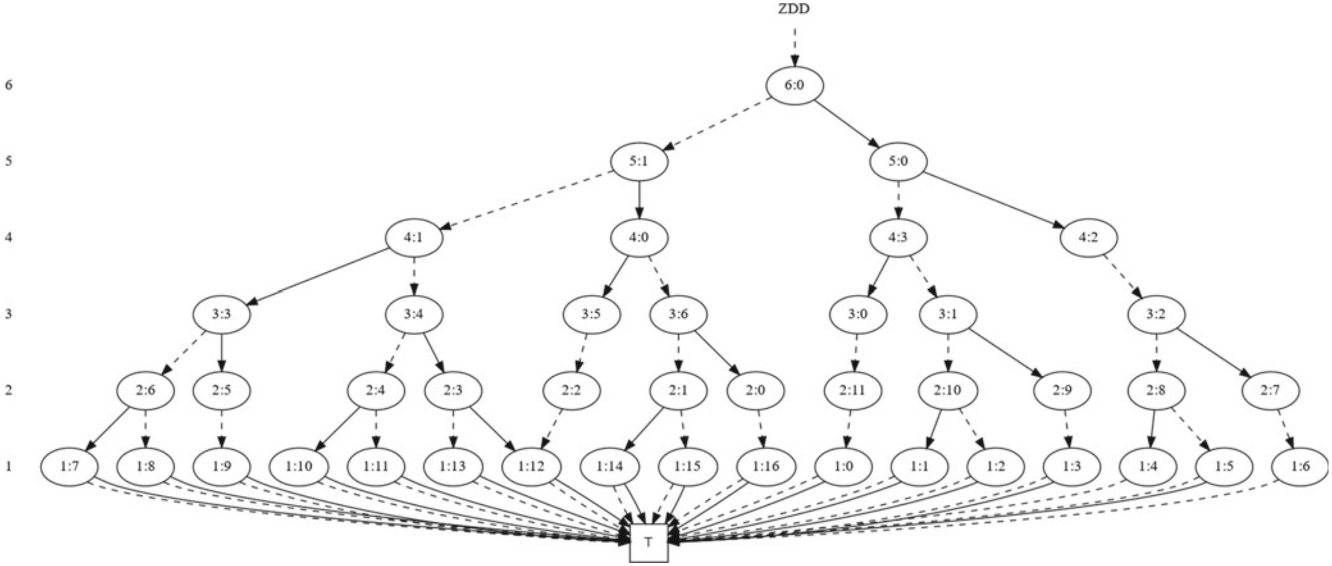


Fig. 2 The rotated diagram for the knapsack problem: the knapsack capacity is 190 for 6 distinct items weighted by (56, 59, 80, 64, 75, 17) and valued by (50, 50, 64, 46, 50, 5). Taking items 1, 2, and 5 yields a weight of 190 and the maximum value of 150

with $F_0 = F_{|E|} = \emptyset$. If a vertex $v \in F_{j-1}$, the array `n.deg[v]` is stored into node n with label e_j. Pruning is done if the set degree constraint is not satisfied.

In brief, the frontier-based search generates a zero-suppressed binary decision diagram through node sharing and pruning. The algorithm concurrently inputs the necessary information in each node of the diagram, enabling the possibility of subgraph representation. Do note, nonetheless, that more information than what is in `n.deg` may be needed and controlled for other types of subgraphs. A complete treatment of the approach is in [18].

5 Methods

Test grid configurations and chosen benchmark graphs of increasing size are converted into an edge list. The edge list would then be divided into two sublists—one for the required edges and another for the nonrequired edges. The first rows of the sublist containing required edges are shown as example.

```
3 4 1
11 12 1
13 14 1
```

A row represents an edge and its properties. Since an edge is defined by a set of two vertices, the first two numbers are the start and end vertices forming the edge. The cost of the edge, in most cases a distance or time measure, is specified by the third number. Visualization is done with the help of an online tool[1] for graphing.

The proposed decision-diagram-based approach in solving the undirected rural postman problem is outlined in Algorithm 1. The explanation follows.

The class `readGraph` begins `zddURPP` by inserting each edge into the graph twice. This workaround is imperative to allow `frontierBasedSearch` to use an edge at most two times. Every required edge is assigned an integer from 0 to $|R| - 1$ in `r`. The two copies of an edge is given the same identification number. Edges that are not required are assigned the number -1.

Following the steps supra, the maximum degree for a vertex in the graph is stored through `getMaxDegree`. This piece of information is required primarily by `degreeConstraint`, which prescribes the vertex degrees during the walk enumeration using `frontierBasedSearch`.

Whether or not each required edge labeled from 0 to $|R| - 1$ is used in the solution is kept track with `Required`. An important procedure in `Required` is `getChild`, seen in Algorithm 2. First, the required edges, including the nonrequired edges, are ensured to be sorted in order of requirement. Upon execution, the constraint sets a flag to true whenever an edge is used. The flag is checked whenever the current level being processed differs from the previous level based on the edge identification

[1] https://csacademy.com/app/graph_editor/.

Algorithm 1 ZDDURPP

```
$G \Leftarrow readGraph()$
$d \Leftarrow$ edge distance function
$r \Leftarrow$ edge requirement assignment

$mvd \Leftarrow getMaxDegree(G)$        ▷ Store the maximum degree of a vertex in G
$dc \Leftarrow degreeConstraint(G)$     ▷ Create a record of degree constraints for vertices in G
$fbs \Leftarrow frontierBasedSearch(G)$ ▷ Create a diagram of all walks in G
$req \Leftarrow Required(G)$            ▷ Create a diagram constraint to use all required edges

for all $v \in G.V$ do
    $dc[v] = [0, 2, 4, \ldots, mvd]$
end for

$zdd \Leftarrow dc \cap fbs \cap req$
$ans \Leftarrow zdd.evaluateMinResult()$
```

Algorithm 2 GETCHILD for REQUIRED

```
procedure $getChild(flag, level, value)$

    if $r[level] \neq r[level - 1]$ then
        if $flag = 0$ then
            return 0
        end if
        $flag = 0$
    end if

    if $value = 1$ then
        $flag = 1$
    end if

    $level \Leftarrow level - 1$
    if $level = 0$ then
        return -1
    end if
    return $level$

end procedure
```

number. If the flag is true then it is reset to false and the the process continues; otherwise, the concerned branch of the zero-suppressed binary decision diagram is cut off. Edges with labels of -1 are skipped.

The intersection of the `degreeConstraint`, `frontierBasedSearch`, and `Required` zero-suppressed binary decision diagrams is taken afterward. Ultimately, the overall optimal walk for the graph is acquired following [19].

Implementation was done in the C++ language. The compiler used was g++ version 7.5.0. The TdZdd[2] library with documentation in [16] was employed for

[2] https://github.com/kunisura/TdZdd.

Table 1 The specifications of the machine used in experimentation

Operating system	Ubuntu 18.04.3 Long Term Support (Bionic Beaver)
Processor	Intel® Core$^{\text{TM}}$ i9-9920X at 3.50 GHz
Graphics card	NVIDIA® GeForce® RTX2080 Ti/PCIe/SSE2
Memory	128 GB

diagram manipulation. Moreover, the framework utilized is based on ZDDLines,[3] the program used in [27]. Refer to the URPDD[4] repository for the complete code. Table 1 summarizes the machine specifications.

6 Results

The experiments are discussed in three parts. Results from preliminary tests done on a sample grid is detailed in the first subsection. The second subsection presents the outcome from solving benchmark instances in literature [4]. The third contains some notes on the performance of the zero-suppressed binary decision diagram method.

6.1 Tests on a Grid Network

A test grid of 3 edges by 3 edges is used as setting. The total number of vertices is 16 and the total number of edges is 24. All edges are set to be of cost 1 unit. Varying configurations of the undirected rural postman problem were then generated, with $3 \leq |R| \leq 13$. The required edges were chosen arbitrarily for each problem.

The results from 8 problems are arranged by the number of required edges in Table 2. Every problem is given an identification number i for reference. To recapitulate, $|E|$ is the number of edges, $|R|$ is the number of required edges, and $|V|$ is the number of vertices. The number of nodes and elements in the resulting zero-suppressed binary decision diagram are denoted by $n(\mathcal{D})$ and $|\mathcal{D}|$, respectively. The time taken in seconds, the memory consumed in MB, and the cost z of the minimal solution is shown in the table.

As seen, the undirected rural postman problem is solved by the approach in reasonable time. Diagrams with tens of thousands of nodes representing billions of feasible solutions is constructed in less than a second on average. In addition, the minimum costs are attained exactly by the routine.

[3] https://github.com/renzopereztan/ZDDLines.

[4] https://github.com/renzopereztan/URPDD.

Table 2 Results of the experiment on the generated test grid networks

i	$\|E\|$	$\|R\|$	$\|V\|$	Time (s)	Space (MB)	$n(\mathcal{D})$	$\|\mathcal{D}\|$	z
1	24	3	16	0.07	5	1.93E+04	2.99E+09	12
3	24	4	16	0.04	4	6.43E+03	2.25E+09	12
2	24	5	16	0.15	9	3.06E+04	1.72E+09	14
5	24	7	16	0.05	5	4.77E+03	1.08E+09	12
4	24	7	16	0.29	16	3.32E+04	1.00E+09	14
7	24	8	16	0.43	23	2.89E+04	8.52E+08	14
8	24	9	16	0.25	15	1.00E+04	6.35E+08	16
6	24	13	16	1.52	66	1.85E+04	2.69E+08	20

6.2 Instances from Literature

Benchmark instances from [4] are then chosen. A selection of 14 problems with $7 \leq |V| \leq 26$ and $10 \leq |E| \leq 47$ is solved. The costs of the edges differ per instance. The number of required edges in a problem ranges from 4 to 24. The required subgraph is determined in accord with the reference text.

Following the same format as the previous table is Table 3, in which the results are found. In the experiment, the generation of a diagram with hundreds of nodes representing thousands of solutions takes a fraction of a second whereas generating a diagram with millions of nodes representing tens of trillions of solutions would take around an hour. Given the difficulty of the task, such consumption of time is justified for decision-diagram-based enumeration [25, 27].

With regard to correctness, the algorithm gives the exact answers without fail. It outperforms the heuristics compiled in [24]. Against exact solutions, the enumerative solution demands more time in calculation. The advantages of the latter, however, are plenty. Apart from the obvious listing of all feasible solutions and drawing of the minimal walk, these include being able to produce the k solutions of lowest or highest cost, to find the mean and variance of all feasible solutions, to filter solutions based on some desired criterion, and others.

6.3 Notes on Performance

To gain deeper insight into the potentiality of the method, some observations are highlighted. One zeroes in on the results for two problem pairs—the fourth and fifth grid network configurations and the fourth and fifth problems from literature.

1. Given the same graph, more required edges means less decisions for the algorithm to make. There are three possibilities for an edge from a zero-suppressed binary decision diagram perspective. These are take the edge, take the edge

Table 3 Results of the experiment on standard instances from literature

i	$\|E\|$	$\|R\|$	$\|V\|$	Time (s)	Space (MB)	$n(\mathcal{D})$	$\|\mathcal{D}\|$	z
13	10	4	7	0.01	4	2.30E+02	3.53E+03	35
1	13	7	11	0.02	4	2.56E+02	5.62E+03	76
11	14	7	9	0.04	4	8.45E+02	1.39E+05	23
12	18	5	7	0.08	6	1.04E+04	5.96E+07	19
10	20	10	12	0.42	24	1.47E+04	2.44E+07	80
9	26	14	14	0.54	39	6.97E+03	3.75E+09	83
2	33	12	14	50.12	2230	4.91E+05	1.03E+12	152
5	35	16	20	31.01	1724	1.74E+05	1.13E+13	124
4	35	22	17	12,947.53	71,229	6.73E+05	3.89E+12	84
15	37	19	26	1714.08	106,885	1.09E+06	1.47E+11	441
18	37	16	23	4532.78	126,383	9.26E+06	1.40E+13	146
8	40	24	17	1552.16	114,617	9.48E+05	1.71E+14	122
17	44	17	19	3481.13	126,641	5.71E+06	4.06E+16	112
7	47	24	23	5289.39	126,723	6.85E+05	6.62E+16	130

twice, and do not take the edge. For a required edge, the option of not taking the edge is eliminated, substantially reducing processing.

2. The problem may be thought of as being the task of connecting parts of the required subgraph through choosing nonrequired edges that would serve as links. This means that the more connected the required subgraph is to begin with, the more efficient the algorithm becomes. Figures 3 and 4 provide an example, where the solution to the fifth grid problem is secured faster than the solution to the fourth possibly due to the required subgraph of the fifth being more connected in the first place.
3. Figures 5 and 6 illustrate how the structure of the given graph affect execution drastically. The fifth benchmark instance is solved in 31.01 s; the fourth benchmark instance, having the same number of edges but a different structure, is solved in 1297.53 s. The sparsity of a graph is a huge factor in computation, especially in the case of a solution by enumeration using the zero-suppressed binary decision diagram.

7 Conclusion

The chapter has advanced an enumerative technique for the solution of the undirected rural postman problem based on the zero-suppressed binary decision diagram. Exact solutions to a diverse set of problem instances were reached in time well-justified

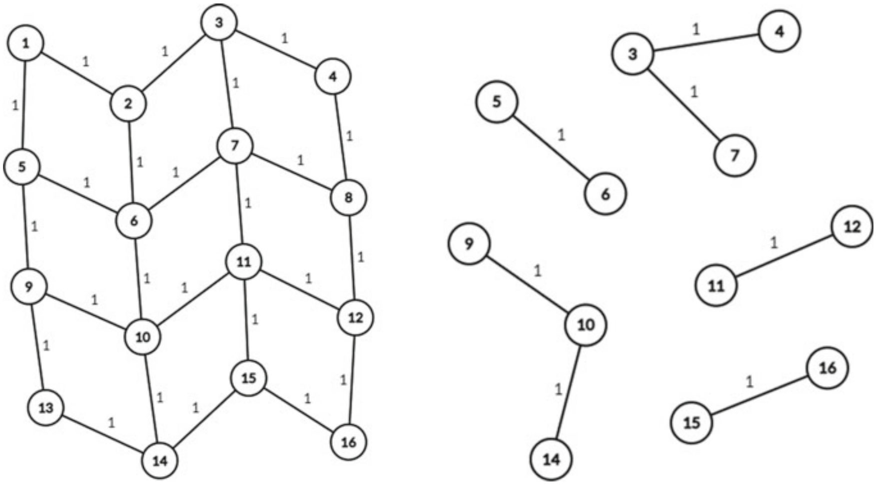


Fig. 3 The given graph and the required subgraph for the fourth sample grid

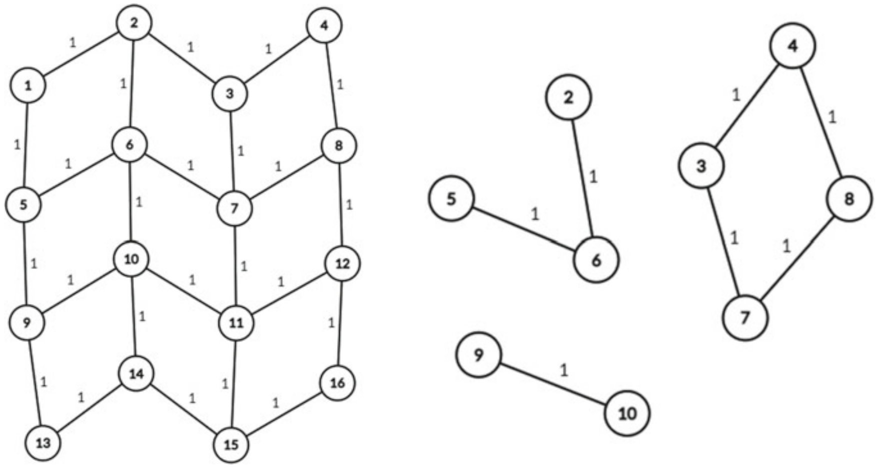


Fig. 4 The given graph and the required subgraph for the fifth sample grid

for methods of enumeration. Observations on performance were pointed out towards the improvement of the algorithm.

For future research, preprocessing may be done to simplify the graphs before diagram construction. Another possible plan of action to reduce computation time is parallelization. As prospects for comparison are scarce, other classes of binary decision diagrams may also be considered so that the proposed approach may be set side by side with other decision-diagram-based solutions.

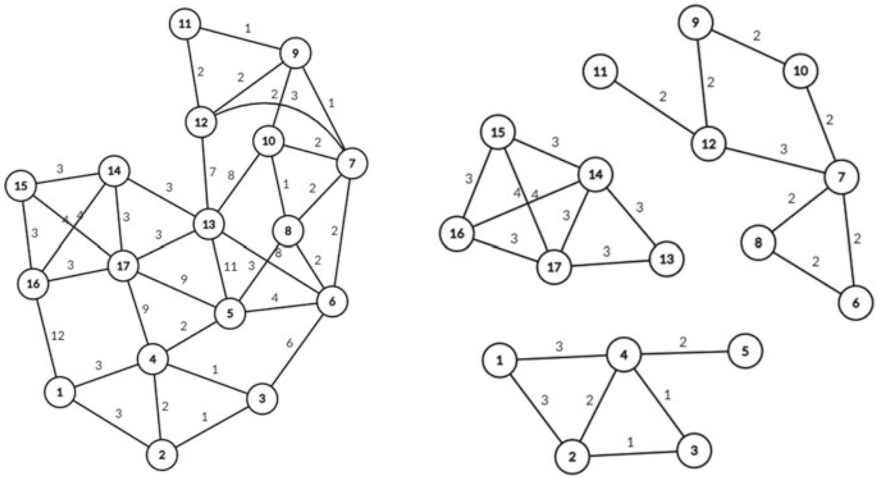


Fig. 5 The given graph and the required subgraph for the fourth benchmark problem

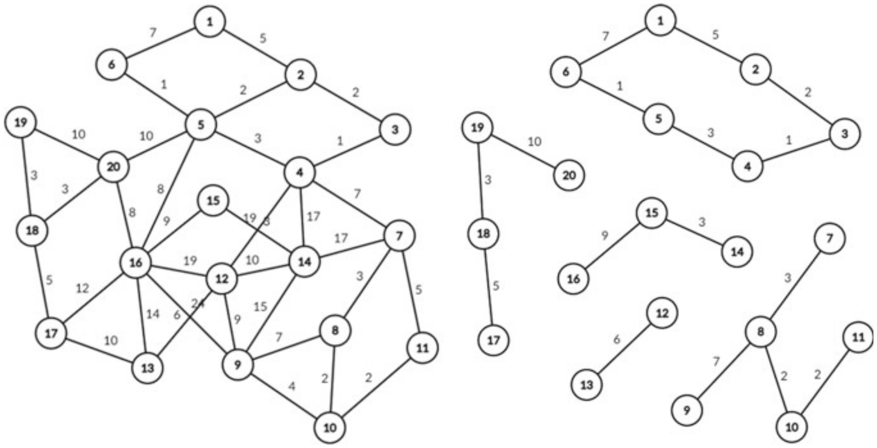


Fig. 6 The given graph and the required subgraph for the fifth benchmark problem

Acknowledgements The research is supported by the Japan Society for the Promotion of Science through the Grants-in-Aid for Scientific Research Program (KAKENHI 18K19821).
Valuable advice was received from Jun Kawahara. J. Kawahara is with the Kyoto University, Japan and the Nara Institute of Science and Technology, Japan. Further, Miguel Zenon Nicanor L. Saavedra has provided assistance in implementation. M. Z. N. L. Saavedra is with the Trainocate Group, Philippines and the Ateneo de Manila University, Philippines.

References

1. Angel, R., Caulde, W., Noonan, R., Whinston, A.: Computer-assisted school bus scheduling. Management Science pp. 279–288 (1972)
2. L. Bodin, S. Kursh, A computer-assisted system for the routing and scheduling of street sweepers. Operations Research **26**(4), 528–537 (1978)
3. Braca, J., Bramel, J., Posner, B., Simchi-Levi, D.: A computerized approach to the new york city school bus routing project. Columbia University Working Paper (1993)
4. Christofides, N., Campos, V., Corberan, A., Mota, E.: An algorithm for the rural postman problem. Imperial College London Technical Report IC.OR.81.5 (1981)
5. Corberan, A., Sanchis, J.: A polyhedral approach for the rural postman problem. European Journal of Operations Research Vol. 79 Iss. 1 pp. 95–114 (1994)
6. de Cordoba, P.F., Garcia-Raffi, L., Sanchis, J.: A heuristic algorithm based on monte carlo methods for the rural postman problem. Computers and Operations Research Vol. 25 Iss. 12 pp. 1097–1106 (1998)
7. J. Desrosiers, J. Ferland, J. Rousseau, G. Lapalme, L. Chapleau, TRANSCOL: A multi-period school bus routing and scheduling system. Studies in the Management Sciences **22**, 47–71 (1986)
8. J. Edmonds, E. Johnson, Matching, euler tours, and the chinese postman. Mathematical Programming **5**(1), 88–124 (1973)
9. H. Eiselt, M. Gendreau, G. Laporte, Arc routing problems, part ii: The rural postman problem. Operations Research **43**(3), 399–414 (1995)
10. E. Fernandez, O. Meza, R. Garfinkel, M. Ortega, On the undirected rural postman problem: Tight bounds based on a new formulation. Operations Research **51**(2), 281–291 (2003)
11. G. Frederickson, M. Hecht, C. Kim, Approximation algorithms for some routing problems. SIAM Journal on Computing **7**(2), 178–193 (1978)
12. Ghiani, G., Laporte, G.: Arc Routing: Problems, Methods, and Applications, chap. The Undirected Rural Postman Problem, pp. 85–99. Society for Industrial and Applied Mathematics (2015)
13. E. Haslam, J. Wright, Applications of routing technologies to rural snow and ice control. Transportation Research Record No. **1304**, 202–211 (1991)
14. Hertz, A.: Graph Theory, Combinatorics, and Algorithms, chap. Recent Trends in Arc Routing, pp. 215–236. Springer (2006)
15. A. Hertz, G. Laporte, P. Hugo, Improvement procedures for the undirected rural postman problem. INFORMS Journal on Computing **11**(1), 53–62 (1999)
16. Iwashita, H., Minato, S.: Efficient top-down ZDD construction techniques using recursive specifications. Hokkaido University Division of Computer Science TCS Technical Report TCS-TR-A-13-69 (2013)
17. Iwashita, H., Nakazawa, Y., Kawahara, J., Uno, T., Minato, S.: ZDD-based computation of the number of paths in a graph. Hokkaido University Division of Computer Science TCS Technical Report TCS-TR-A-12-60 (2012)
18. Kawahara, J., Inoue, T., Iwashita, H., Minato, S.: Frontier-based search for enumerating all constrained subgraphs with compressed representation. IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Sciences Vol. E100-A No. 9 pp. 1773–1784 (2017)
19. Knuth, D.: The Art of Computer Programming Vol. 4 Fasc. 1. Addison-Wesley (2009)
20. Kreher, D., Stinson, D.: Combinatorial Algorithms: Generation, Enumeration, and Search. Chemical Rubber Company Press (2009)
21. J. Lenstra, A.R. Kan, On general routing problems. Networks **6**(3), 273–280 (1976)
22. S. Minato, Zero-suppressed BDDs and their applications. International Journal on Software Tools for Technology Transfer **3**(2), 156–170 (2001)
23. Minato, S.: Zero-suppressed BDDs for set manipulation in combinatorial problems. Proceedings of the 30th International Design Automation Conference pp. 272–277 (Dallas, United States of America, 1993)

24. Moreira, M., Ferreira, J.: A genetic algorithm for the undirected rural postman problem (2010)
25. D. Morrison, E. Sewell, S. Jacobson, Solving the pricing problem in a branch-and-price algorithm for graph coloring using zero-suppressed binary decision diagrams. INFORMS Journal on Computing **28**(1), 67–82 (2016)
26. Orloff, C.: A fundamental problem in vehicle routing. Networks Vol. 4 Iss. 1 pp. 35–64 (1974)
27. Tan, R., Kawahara, J., Garciano, A., Sin, I.: A zero-suppressed binary decision diagram approach for constrained path enumeration. Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2019 pp. 132–136 (3-5 July 2019, London, United Kingdom)
28. T. Toda, T. Saitoh, H. Iwashita, J. Kawahara, S. Minato, ZDDs and enumeration problems: State-of-the-art techniques and programming tool. Computer Software **34**(3), 97–120 (2017)

Building an Academic Identity: Benefits and Drawbacks in Multicultural Engineering Classrooms

Simona Vasilache

Abstract In our globalized world, multicultural instruction poses various challenges to educators and students alike. While these challenges have been traditionally recognized in the field of social sciences, they have been neglected by the natural sciences. Our study aims to illustrate the importance of cultural differences in science and engineering education. Based on data gathered in two multicultural environments, at two academic institutions from two different countries, we highlight the students' perceptions of the cultural differences, how they help build an academic identity and how they influence the teaching and learning process.

Keywords Cultural differences · Education · Engineering · Multiculturalism · Student diversity · Student perception

1 Introduction

We live in a globalized, internationalized, multicultural world; these are words that we hear almost daily, and we may miss their various implications. We come upon multicultural environments everywhere; the academic world is one such environment and, in it, multiculturalism brings along a multitude of challenges, for both educators and students.

One of earliest definition, given by one of the pioneers of multicultural education, James Banks, in 1989, states that multiculturalism is "a philosophical position and movement that deems that the gender, ethnic, racial, and cultural diversity of a pluralistic society should be reflected in all of the institutionalized structures of educational institutions, including the staff, the norms, the values, the curriculum, and the student body" [1]. In the past three decades, along with a rapid expansion, multicultural societies have suffered various transformations. As Xie et al. stated in [2], "while globalization has resulted in shorter distances between individuals,

S. Vasilache (✉)
Graduate School of Systems and Information Engineering, University of Tsukuba, Tennodai 1-1-1, Tsukuba, Japan
e-mail: simona@cs.tsukuba.ac.jp

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_10

cross-cultural problems arise in many aspects, especially communication conflicts caused by cultural diversity".

In the field of education, cultural differences intrinsic to any international classroom bring along a series of challenges that must be overcome by educators and pupils alike. According to Parrish and Linder-VanBerschot, "the growing multicultural nature of education and training environments makes it critical that instructors and instructional designers [...] develop skills to deliver culturally sensitive and culturally adaptive instruction" [3]. The internationalization of higher education has greatly transformed university classrooms and instructors must be "well equipped to teach in such culturally diverse context while sustaining the goals of internationalization" [4].

As shown in [5], "mastering and implementing various teaching styles is the ideal way to match up the needs of diverse learners, the variety of content knowledge, and educational goals". Culturally responsive teaching, first defined in 1994 by Ladson-Billings as "a pedagogy that recognizes the importance of including students' cultural references in all aspects of learning" [6], is gaining ground and is increasingly considered by educators in all kinds of academic institutions. Culturally responsive education implies changes to the teaching process that include "varying teaching styles, employing flexible grouping, and collaborating with students, in order to create a more cooperative learning climate" [7].

Traditionally, culturally differences were recognized in social sciences education; natural sciences tend to be viewed as less culturally dependent. Recently, however, a large body of research has been dedicated to studies on culturally responsive teaching in science, math etc. [8]. We believe that they deserve equal attention in terms of culturally responsive teaching; our study intends to bring out its importance, based on data gathered from two groups of international students, in two different countries, by highlighting their perceptions of multiculturality and diversity in the classroom.

The remainder of our paper is organized as follows: Sect. 2 describes the research setting and the method used, while Sect. 3 presents the results and discusses their implications. Section 4 concludes our paper and points out directions for future research.

2 Research Setting and Method

The results presented in this study are based on two sets of data. One set of data was collected by the author during a visit to the Faculty of Engineering in Foreign Languages (FILS) at the Politehnica University of Bucharest in Romania, in November 2018 (henceforth called "UPB data"). Together with the local English instructor, the author attended and peer-taught a practical course on the topic of "Teaching strategies in multicultural environments". The goals of the course were identifying cultural aspects of education, in relation to teaching style and course structure, as well as finding "the best way" to teach in multicultural settings [9].

The course participants, coming from a total of 16 different countries, were international and local Romanian students enrolled in the Information Technology course of FILS, with English as the main language of instruction. The students' contribution was two-fold. At the beginning of the course, they filled in an online questionnaire online (157 responses were collected); next, they participated in two focus groups encouraging class reflection on the topic of multicultural classrooms, seen from the perspective of the students, as well as that of the instructors.

The second set of data was collected through two questionnaires, administered in February 2019 and December 2019, respectively, during the so-called practical part of the computer literacy classes, taught by the author and held in the undergraduate English programs at the University of Tsukuba in Japan (throughout this paper, we will refer to this data set as "UT data"). Until spring 2019, this course (mandatory for all freshmen, regardless of their field of study) was called "Information Literacy (Practice)"; starting with April 2019, the course was renamed "Information Literacy (Exercises)". The participants were 56 freshmen, coming from 15 different countries, studying in English in four different undergraduate programs (social and international studies, life and environmental sciences, global issues, interdisciplinary engineering). These programs are aimed at international students, who can obtain a bachelor's degree through courses held in English only.

One common feature of the two sets of participants is that they consist of students accustomed to studying in a multicultural environment; by definition, they are enrolled in programs aimed at international students. Furthermore, whereas the language of instruction is English for all the participants, most of them are non-native English speakers.

3 Results and Discussion

The data collected will be viewed from three different perspectives, as in the following three subsections. First, multicultural instruction in general will be considered; next, teaching and learning aspects in multicultural classrooms will be highlighted, followed by class interactions in multicultural environments. In each subsection, the results of the questionnaires (from both UPB and UT data sets) will be presented first; they will be followed by data extracted from the peer-teaching sessions which took place at UPB. The focus group discussions were recorded, and the scripts of the recordings were analyzed; in them, the students expressed various thoughts on multicultural classrooms and cultural differences. Partial results of the questionnaires were presented in our previous work (see [10]).

3.1 Multicultural Classrooms

In the first part of the questionnaire, the participants were asked about their opinions on advantages and disadvantages of being in an international/multicultural classroom (multiple answers were possible).

When it comes to advantages, most participants (31.56%) chose "learning about new cultures/broadening one's horizon", followed by "becoming culturally flexible/aware" (27.81%) and "making new friends" (26.23%). Moreover, 11.44% of all participants chose "questioning one's own cultural values"; less than 2% (i.e. 1.97%) of participants found no advantage at all. Two more advantages were mentioned by the participants: opportunity to practice and learn other languages (two students) and "social networking and building beneficial relationships" (one student). The combined results from the two sets of data appear in Fig. 1.

With regard to disadvantages, as shown in Fig. 2, 35.77% of all participants chose the extra effort needed for social interactions. Almost one third of the students (i.e. 32.31%) found no disadvantage at all. Cultural clashes were chosen by 19.23% of participants, whereas 9.62% chose uncertainty about own cultural values. Other answers provided by the participants included feeling "left out", language barrier "issues and misunderstandings", being "offensive or ignorant" with regard to other cultures, along with difficulty of interacting with people from other cultures, "since their views, style of life, values and so on are more or less different". In short, approximately two thirds of the participants recognized the social and cultural issues that may arise in a multicultural group. While this may pose some difficulties for the instructor conducting the multicultural classes, it is important to note the students' awareness of these cultural differences.

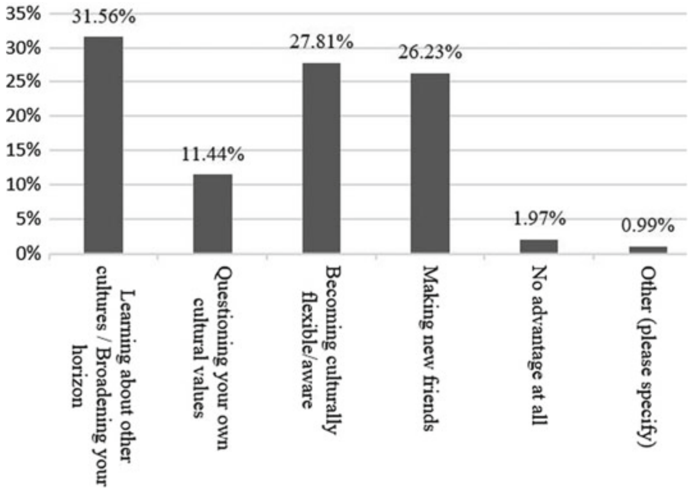


Fig. 1 Advantages of being in an international/multicultural classroom

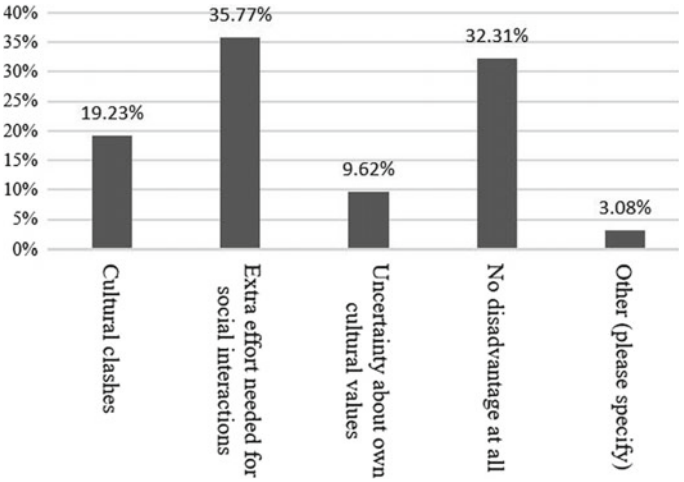


Fig. 2 Disadvantages of being in an international/multicultural classroom

During the group discussions (part of the UPB data), students underlined the fact that "*a lot of different cultures […] can be good actually; at the same time, it can be bad*". One participant mentioned cultural clashes, stating that they are "*not happening that often, but you need [to] put extra effort for social interactions because there's a language barrier*". The same student added: "*we're not used to speaking in other languages apart from our mother tongue, so I think this is the only disadvantage*". One other participant mentioned the friendliness in class, regardless of the cultural background, while contrasting this with the outside world, where historical/political etc. issues may perpetuate animosities between certain larger social groups, even nations. Moreover, an example was given of a student from country XXX (where the level of high-school mathematics is supposedly low), who came to study at UPB and who could not keep up with some of the courses. This student seemed to have asked for help repeatedly and during a rather long period of time; while it was agreed that help should be given when asked for, taking advantage of someone and using up a lot of their time was found to be unacceptable. It is possible that cultural differences were implied when this type of situations arise.

One other participant stated the following: "*I don't think you have a disadvantage; you always know other people, other cultures, other type of studies*". Moreover, in a multicultural classroom, students have the opportunity to refine their English skills (i.e. speaking and listening), as well. According to different students, meeting people from other countries "*makes you more open minded*", makes you "*enriched*", and "*perhaps a little more tolerant, adaptable*" or "*you may change your point of view on something you are <fixed on>*".

3.2 Teaching and Learning in Multicultural Classrooms

Through the questionnaires, the participants in both groups were asked if they believe that, based on the culture they belong to, students approach learning and teaching methods differently. As shown in the chart in Fig. 3, more than three quarters of the participants (77.36% in total) chose either “yes, clearly” (32.55%) or “yes, I think so” (44.81%); the remaining 22.64% respondents chose “no, I don’t think so” (20.28%) or “not at all” (2.36%). With regard to learning styles of students from different cultures, the UPB student group expressed thoughts like: “*I wouldn’t really say there’s a particular difference, but rather a generality because being raised in different cultures means they have a different view over [...] the world*”, or “*in general, each culture sees a difference in the world*”.

When asked if they believe that their instructors should change/adapt their teaching style depending on the cultural background of the class participants, as Fig. 4 shows, a total of 43.19% either strongly agree or agree, whereas 42.25% neither agree nor disagree; a total of 14.55% disagree or strongly disagree. It is interesting to observe that, while being aware of cultural differences in a multicultural classroom, the students do not seem to have a particular expectation of the instructor changing or adapting their teaching style, so as to accommodate the students’ cultural differences. Whatever difficulties they may encounter, the students appear to find it normal for them to overcome these difficulties, without any help from the instructors. We could argue that, while many researchers believe it to be a necessity, culturally responsive teaching [5] is not a common practice yet; it is highly possible that the concept is not even heard of by some students.

During the discussions in the focus groups, the participants were asked how a teacher can address all students from different cultures successfully. In their answers,

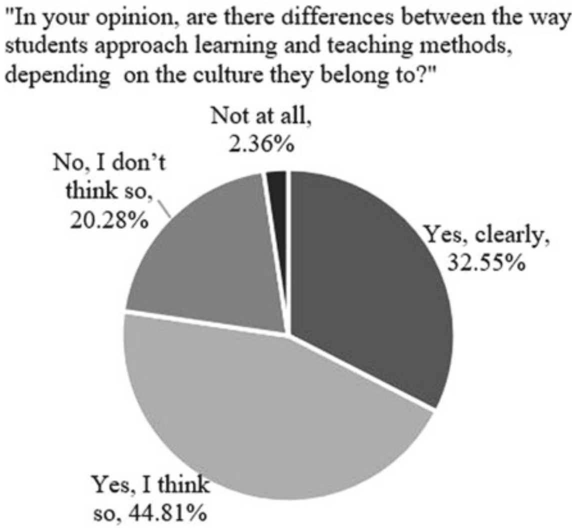


Fig. 3 Perceptions of differences in learning/teaching methods

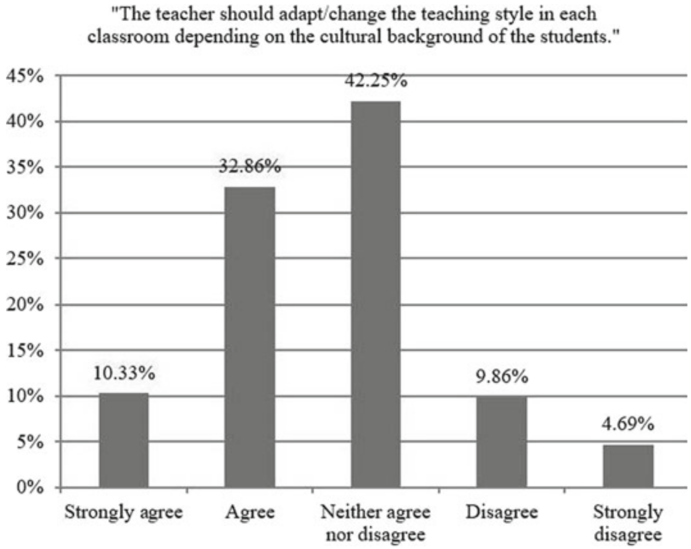


Fig. 4 Perceptions of necessity of adapting teaching style in multicultural classrooms

the students stated that, for instance, in discriminatory situations, when students may "look down" on other students, the teacher should "directly confront the students", by simply saying "*look, I've seen this and it's wrong*". Several students believe cultural differences are difficult to handle, e.g. "*it is impossible to not offend some other people from other culture*", "*some cultures are contradicting; some of them are really contradicting*", "*you should choose the lesser evil*". One participant noted that "*if you try to not offend anybody, you will end up not saying anything*".

As noted in Sect. 1, cultural differences have traditionally been considered in "social sciences"; until recently, the so-called "natural sciences" were not regarded as culturally dependent. Based on the questionnaire responses, our study participants' view on this issue is summarized in Fig. 5. They were presented with the following statement: <*Cultural differences in the classroom are more visible in "social sciences" (e.g. literature, history* etc.*) than they are in "natural sciences" (e.g. mathematics, engineering* etc.)>. Out of all participants, 19.72% strongly agree, whereas 45.07% agree with this statement. A total of 7.98% respondents either disagree or strongly disagree, whereas 27.23% neither agree nor disagree with the given statement.

The group of University of Tsukuba students responded to our questionnaire during their computer literacy related class. For this group only, one question was posed, i.e. whether, in the context of a multicultural classroom, computer literacy is more, less or equally challenging than studying other subjects. As can be observed from the results (shown in Fig. 6), 57.14% participants consider that learning computer literacy is the same as learning any other subject and 23.21% believe that it is easier than learning other subjects. Only 8.93% believe that computer literacy

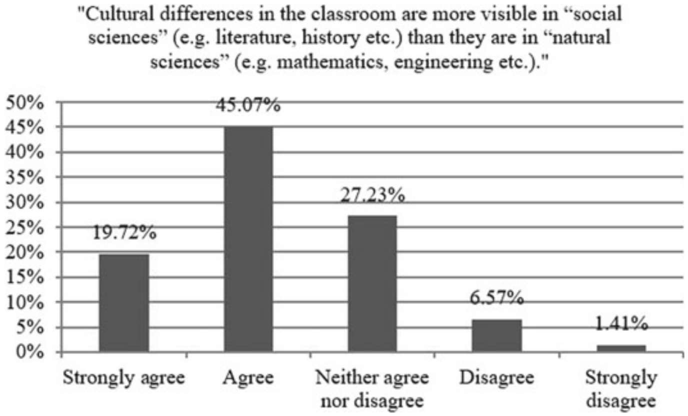


Fig. 5 Cultural differences in "social sciences" versus "natural sciences"

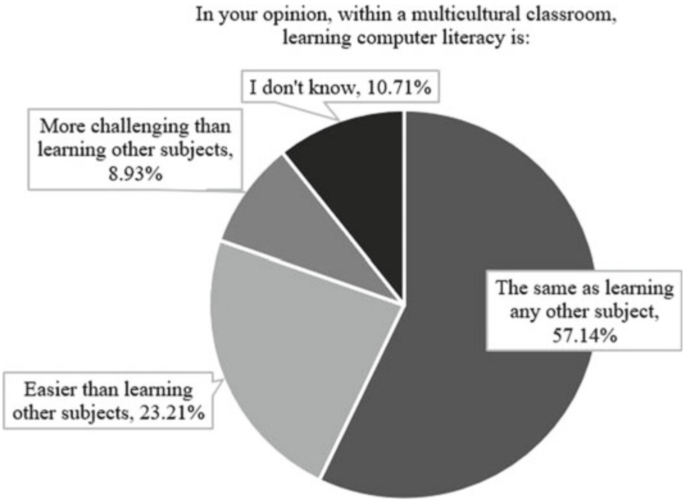


Fig. 6 Comparison between studying computer literacy and other subjects

is more challenging than other subjects, in a multicultural context. (A significant proportion of students, i.e. 10.71%, replied that they do not know.) Civitillo et al. reached a similar conclusion in their work in [8], stressing the "difficulty to implement culturally responsive teaching in subjects like science and math" [11–13], which are "often considered culture-free and objective by nature" [14].

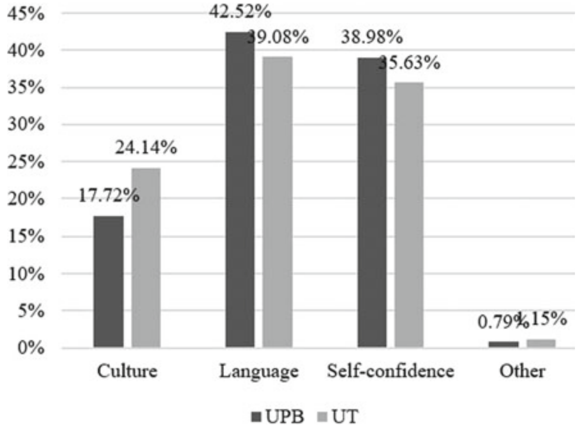


Fig. 7 Most difficult issue during class discussions

3.3 *Multicultural Class Interactions*

In multicultural classrooms, the interactions between students are highly significant. Our questionnaire gathered data with regard to the most difficult issue to deal with during class discussions. The participants had the following choices: cultural differences (people from different cultures have different styles of arguing/discussing), language issues (some people speak the language better/worse than the others), self-confidence issues (some people are too shy, while others are the exact opposite) and other. Figure 7 illustrates the results from the two sets of participants. We can observe that language ranks highest for both the UPB and the UT groups (42.52% and 39.08, respectively), followed by self-confidence (38.98 and 35.63, respectively). Interestingly, the least difficult issue out of the three suggested is "cultural differences" : 17.72% for the UPB group and 24.14% for the UT group.

During the focus group discussions, language was mentioned as the most prevalent issue to deal with, along with "personality".

When it comes to working with colleagues belonging to different cultural backgrounds, we wanted to find out what the students consider as the source of most difficulties. This issue was included in the questionnaire administered to the UT students in particular, because one of the assignments for the computer literacy class is based on a group project, in which students (from various countries) work together on preparing and holding a presentation in front of the whole class. The responses were as follows: almost half (46.43%) chose "different personalities", 30.36% chose "variations in language skills" and the remaining 23.21% chose "different work styles".

As a partial measure of how international students perceive difficulties in multicultural instruction, away from their home countries, the UT group of students were asked how likely it was for them to recommend studying abroad to a friend, on a scale of 0–10. While we agree that there may be some controversies regarding its use in academia, we calculated the Net Promoter Score [15] for this aspect and we found that it is 37.5 (half of the participants were promoters, i.e. 50% scoring 9–10, whereas 12.5% of the participants were detractors, scoring between 0 and 6). This is an inspiring result, showing that students focus on the benefits of multicultural instructions, rather than the drawbacks, and they would consequently encourage their friends to study abroad.

4 Conclusions and Future Work

Based on a study performed in two different academic institutions, our paper identified engineering students' perceptions of studying natural sciences in multicultural/international environments. We concluded that, while they are acutely aware of cultural differences in the classroom, the students have no expectations from their instructors to adapt the teaching style based on the different cultural backgrounds of the class participants. Moreover, our study highlighted the need for a culturally adaptive teaching in natural sciences, not only in the traditional social sciences. In our future work, we intend to collect more data and to design a culturally responsive teaching strategy tailored to engineering students.

Acknowledgements This work was partly supported by JSPS KAKENHI 18K02953 [Grant-in-Aid for Scientific Research (C)]. The author would like to thank Dr. Yolanda Catelly for her valuable and constructive feedback and for the support in providing the setting needed to collect part of the data used in this study.

References

1. J.A. Banks, Approaches to multicultural curriculum reform. Trotter Rev. **3**(3), Article 3 (1989)
2. A. Xie, P.P. Rau, Y. Tseng, H. Su, C. Zhao, Cross-cultural influence on communication effectiveness and user interface design. Int. J. Intercultural Relat. **33**(1), 11–20 (2009)
3. P. Parrish, J. Linder-VanBerschot, Cultural dimensions of learning: addressing the challenges of multicultural instruction. Int. Rev. Res. Open Distrib. Learn. **11**(2), 1–19 (2010)
4. R. Heringer, The pertinence of a culturally relevant pedagogy in internationalized higher education. Int. Educ. Stud. **12**(1), 1–9 (2019)
5. H.Z. Zeng, Differences between student teachers' implementation and perceptions of teaching styles. Phys. Educ. **73**(2), 285 (2016)
6. G. Ladson-Billings, *The dreamkeepers: successful teachers of African American children* (Jossey-Bass, San Francisco, 1994)
7. M. Ortiz, Culturally responsive multicultural education. St. John Fisher College, School of Education (2012), p. 8
8. S. Civitillo, L.P. Juang, M. Badra, M.K. Schachner, The interplay between culturally responsive teaching, cultural diversity beliefs, and self-reflection: a multiple case study. Teach. Teach. Educ. **77**, 341–351 (2019)
9. The Department of Communication in Modern Languages, Faculty of Engineering in Foreign Languages, Politehnica University of Bucharest, Romania, Teaching Strategies in Multicultural Environments. Available https://dclm.pub.ro/?p=1995&lang=en

10. S. Vasilache, Science language: universal or culturally dependent? Multicultural instruction and engineering students' perceptions, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, London, UK, 3–5 July 2019, pp. 121–125
11. E.P. Bonner, T.L. Adams, Culturally responsive teaching in the context of mathematics: a grounded theory case study. J. Math. Teach. Educ. **15**, 25–38 (2012)
12. K.J. Debnam, E.T. Pas, J. Bottiani, A.H. Cash, C.P. Bradshaw, An examination of the association between observed and self-reported culturally proficient teaching practices. Psychol. Schools **52**, 533–548 (2015)
13. L. Shumate, G.D. Campbell-Whatley, Y.Y. Lo, Infusing culturally responsive instruction to improve mathematics performance of Latino students with specific learning disabilities. Exceptionality **20**, 39–57 (2012)
14. J.S. Matthews, F. López, Speaking their language: the role of cultural content integration and heritage language for academic achievement among Latino children. Contemp. Educ. Psychol. **57**, 72–86 (2018)
15. F.F. Reichheld, The one number you need to grow. Harvard Bus. Rev. **81**(12), 46–55 (2003)

A Passive Acoustic Method for Detection of Gas Bubbles Size Distribution in Water

Stepan A. Gavrilev, Julia M. Tyurina, and Mikhail V. Ivanov

Abstract Currently, much attention is paid to the dispersed composition of the gas phase in a liquid medium, since this is a key parameter in the design and calculation of most hydromechanical systems. In this work existing methods for determining the dispersed composition of the gas phase in a liquid medium are reviewed. These methods are of three types: optical, photometric and acoustic. In this work we proposed a passive acoustic method for determining the distribution of air bubbles in water. Unlike other passive acoustic methods, its scope is not limited to the nature of the distribution. The method was tested on an experimental setup for determining the distribution of air bubbles in a cubic tank filled with water. The measured bubble noise spectrum was recalculated into a function of the density distribution of bubble size by using the mathematical model described in this work. The correctness of the model was demonstrated by comparing the results with a photometric method based on shadow analysis.

Keywords Acoustic bubble sizing · Acoustic emission spectra · Air bubble · Bubble size distribution · Passive acoustic method · Shadow analyzing

S. A. Gavrilev (✉)
Bauman Moscow State Technical University, Novodmitrovskaya, 2k6-343, 127015 Moscow, Russian Federation
e-mail: stepan.tab92@gmail.com

J. M. Tyurina
Bauman Moscow State Technical University, Baumanskaya, 35/1-14, 105005 Moscow, Russian Federation
e-mail: j.sforcia@gmail.com

M. V. Ivanov
Bauman Moscow State Technical University, Lyotnaya, 21/1-10, 141018 Mytyschy, Moscow Region, Russian Federation
e-mail: mivanov2005@mail.ru

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_11

1 Introduction

As of today, we have the growing number of industrial technologies based on the use of hydromechanical systems containing substances in various phase states, primarily gas and liquid. The dynamics of processes occurring in such systems over the years has been the subject of many studies [1–3]. In such works much attention is paid to the dispersed composition of the gas phase in a liquid medium, since this is a key parameter in the design and calculation of most hydromechanical systems [4, 5].

Existing methods for determining the dispersed composition of the gas phase in a liquid medium can be divided into three types: optical, photometric and acoustic.

Optical methods for determining the size of gas bubbles in liquid media are based on the principles of light blocking [6] or reflection in a dark-field [7]. The principle of these methods is that a laser beam is passed through the bubble and the reflected and refracted light from the bubble is measured by a photo detector.

Photometric methods are based on the analysis of images of a gas-liquid mixture. The correctness of the definition of bubbles depends on many factors. First of all, on the quality of images (resolution and contrast), it should be sufficient so that every single bubble can be distinguished. Secondly, on the applied algorithm of image analysis. Basically the shadow analysis technique is used, for example, as in the works [8–10]. The disadvantage of this approach is that for each specific gas-liquid mixture, preliminary complex calibration is required with reference to the conditions of photography (illumination, optical density of the medium, light reflection from various elements, etc.) It is also worth noting that the analysis is very difficult if there are adjacent, overlapping bubbles in the images.

The main disadvantage of optical and photometric methods is that they are applicable only in optically transparent media. As a medium in which monitoring is required is often turbid, the acoustic methods are used.

Most modern acoustic methods are active, i.e. they are based on measuring the attenuation and velocity of an acoustic wave passing through a bubble cloud layer. Examples of the implementation of this approach are described in [11–13]. But it should be considered that such methods have an important feature: the acoustic wave generated to irradiate a bubble cloud can affect the bubbles. Proof of this can be the results of works [14]. The authors of this work conducted a series of experiments, while changing the intensity of the emitted acoustic wave and in each of the experiments obtained different results. Thus, in a number of cases where maintaining modes with a certain bubble size is especially important, the use of active methods is unacceptable. Therefore, it is preferable to use passive methods that do not affect the dynamics of the studied processes.

Passive methods are based on the fact that air bubbles in water are sources of acoustic signals. They emit an acoustic signal due to the variable gas pressure inside the bubble.

In 1933, Marcel Minnaert showed that under adiabatic conditions (heat transfer between the gas of the bubble and the liquid is insignificant), the frequency of the acoustic signal f emitted by the gas bubble depends on its size as follows:

$$f = \frac{1}{2\pi R}\sqrt{\frac{3\gamma P}{\rho}}, \tag{1}$$

where f—frequency of the acoustic wave emitted the bubble; P—absolute fluid pressure; γ—specific heat coefficient of gas, ρ—fluid density; R—bubble radius.

Minnaert's expression also shows a very simple inverse relationship for calculating bubble size from the frequency of the acoustic wave emitted by it. For example, if an air bubble in water emits sound at a frequency of 3000 Hz, then its radius is 1 mm. Later on this dependence was studied and demonstrated by other scientists, for example, in [15–17].

Equation (1) is valid for spherical bubbles. The shape of the bubbles usually changes during floating up. Depending on the size, they can take the form of a sphere, a flattened spheroid of a spherical segment or a mushroom cap. According to the criterion presented in [18], bubbles can be considered spherical if the following condition is met:

$$R < \delta_{\sigma}, \tag{2}$$

where R—bubble radius; δ_{σ}—liquid capillary constant.

According to this criterion, the bubbles in the water have the shape of a sphere if their radius $R < 2.7$ mm.

The most complete description of bubble acoustics was given by Leighton (1994). According to Leighton and Walton [19], the spectrum of sound emitted by bubbles can be used to determine their size. Further, this idea was developed in works [20–24].

But it is worth noting that there is still no universal algorithm for converting the bubble noise spectrum into a size distribution. In most works, only the sizes of single bubbles are determined. For example, in works [22, 23, 25] the authors determine the peak frequency in the spectrum, and the size of the bubbles is found from this frequency through the inverse Minnaert dependence. This approach is applicable only in cases where all the bubbles are approximately the same size and therefore, a peak can be seen in the spectrum.

In works [20, 21, 24] the expressions are presented to determine the acoustic pressure generated by single bubbles, but there are no ratios for calculating the total spectrum.

Of great interest for monitoring are systems with a complex size distribution of bubbles. Such systems are usually characterized by the presence of a large number of bubbles with different sizes, so the technique with finding peak frequencies is not suitable. In this work, we describe a universal algorithm for converting the noise spectrum of bubbles into a size distribution, which is applicable for complex distributions.

2 Theoretical Part

Our task is to determine the size distribution of bubbles. To do this, we introduce the following notation: N—total number of bubbles; $g(r)$—distribution density function, r—bubble radius.

Let us assume that we measured the bubble noise and obtained a spectrum with a constant bandwidth df and the number of bands K. That is, we have K acoustic pressure measurements $P(f_i)$, where: $i = 1, \ldots, K$; f_i—center frequency of ith pass band.

Whereby:

$$f_i = i \cdot df. \tag{3}$$

Then we count how many bubbles "make noise" in the ith frequency band. The boundaries of our frequency band can be written as:

$$f_i - df/2 \quad \text{and} \quad f_i + df/2. \tag{4}$$

According to Minnaert, these boundary frequencies will correspond to bubbles with radii R satisfying the following conditions:

$$M/(f_i - df/2) = R2_i \geq R > R1_i = M/(f_i + df/2), \tag{5}$$

where $M \approx 3000$ [mm Hz]—Minnaert constant.

A diagram of the boundaries of the studied frequency band and the bubble size scale is presented in Fig. 1.

From the distribution density function $g(r)$ and the total number of bubbles N, we can calculate the number of N_i bubbles in the desired interval as follows:

$$N_i = N \cdot (G(R2_i) - G(R1_i)), \tag{6}$$

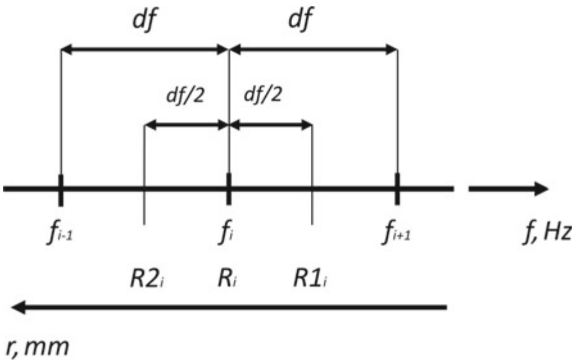


Fig. 1 Scheme of the ith frequency bandwidth

where $G(R) = \int_0^R g(r)dr$—cumulative distribution function.

We number the radii of the found bubbles in an arbitrary order as:

$$R_j, \quad j = 1, \ldots, N_i, \tag{7}$$

We can also relate the pressure in the *i*th frequency band P_i and the number of bubbles N_i. We know that the pressure created by a bubble is proportional to its size. We write it as:

$$p(R) = a \cdot R, \tag{8}$$

where *p*(*R*)—function of acoustic pressure created by a single bubble with radius *R*, *a*—some proportionality coefficient.

From the law of conservation of energy, we can write:

$$P_i^2 = \sum_{j=1}^{N_i} p\left(R_j\right)^2 = \sum_{j=1}^{N_i} \left(a \cdot R_j\right)^2. \tag{9}$$

Since we do not know the exact values of R_j, we make the assumption:

$$P_i^2 = \sum_{j=1}^{N_i} \left(a \cdot R_j\right)^2 \approx N_i \cdot a^2 \cdot R_i^2, \tag{10}$$

where $R_i = M/f_i$—radius corresponding to the center frequency f_i of the *i*th frequency band.

We rewrite the previous expression as follows:

$$N_i \approx P_i^2/(a^2 R_i^2) = P_i^2/(a^2(M/f_i)^2) = P_i^2 f_i^2/(a^2 M^2). \tag{11}$$

Given expression (Eq. 6):

$$N_i = P_i^2 f_i^2/\left(a^2 M^2\right) = N \cdot (G(R2_i) - G(R1_i)). \tag{12}$$

We make one more assumption:

$$G(R2_i) - G(R1_i) = g(R_i) \cdot (R2_i - R1_i). \tag{13}$$

We rewrite expression (Eq. 12):

$$N_i = P_i^2 f_i^2/\left(a^2 M^2\right) = N \cdot g(R_i) \cdot (R2_i - R1_i). \tag{14}$$

From the previous value, we express the desired $g(R_i)$:

$$g(R_i) = P_i^2 f_i^2 / \big(N \cdot a^2 \cdot M^2 \cdot (R2_i - R1_i)\big). \tag{15}$$

To calculate $g(R_i)$ we are missing value $N \cdot a^2$. We find this value by summing expression (Eq. 11) over all K frequency bands:

$$\begin{aligned} \sum_{i=1}^{K} N_i &= \sum_{i=1}^{K} \big(P_i^2 f_i^2 / \big(a^2 M^2\big)\big), \\ N &= \sum_{i=1}^{K} \big(P_i^2 f_i^2 / \big(a^2 M^2\big)\big). \end{aligned} \tag{16}$$

From the previous expression we find the desired value $N \cdot a^2$:

$$N \cdot a^2 = \sum_{i=1}^{K} \big(P_i^2 f_i^2 / M^2\big). \tag{17}$$

Substitute the resulting dependence in Eq. (15):

$$g(R_i) = P_i^2 f_i^2 / \left[(R2_i - R1_i) \sum_{k=1}^{K} (P_k^2 f_k^2) \right]. \tag{18}$$

Taking into account (Eq. 3) and (Eq. 5), we write the final form of the expression for calculating the values of the probability density function $g(r)$:

$$g(R_i) = \frac{P_i^2 f_i^2}{M \cdot df \cdot \left(\sum_{k=1}^{K} P_k^2 f_k^2\right)} \cdot \big(f_i^2 - (df/2)^2\big). \tag{19}$$

And with the assumption that $df^2 \ll f_i^2$, expression (Eq. 19) takes the form:

$$g(R_i) = \frac{P_i^2 f_i^4}{M \cdot df \cdot \left(\sum_{k=1}^{K} P_k^2 f_k^2\right)}. \tag{20}$$

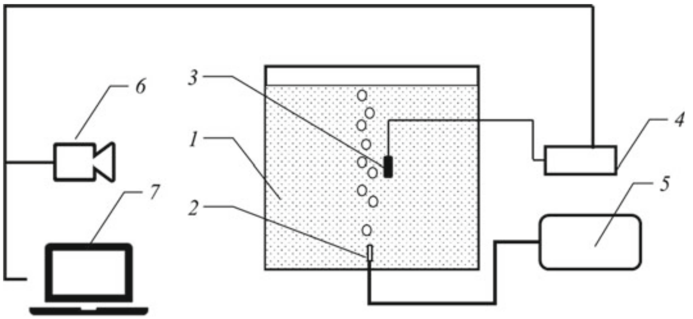


Fig. 2 Scheme of the experimental setup

3 Experimental Part

To verify the proposed theory, it was decided to compare it with another method. The photometric method was chosen as the reference method. For this, an experimental setup was assembled. The scheme of the setup is presented in Fig. 2. It consists of a cubic tank with transparent walls (1), a finely porous aerator (2), a BK type 8103 hydrophone (3), a spectrum analyzer (4), a compressed air cylinder (5), a camera (6) and a PC (7).

Air was passed from the can of compressed air through the aerator and bubbles were thus generated. At this moment, a hydrophone connected to a spectral analyzer measured the spectrum of noise emitted by the bubbles. The hydrophone was located in the center of the reservoir near the stream of bubbles. The bandwidth df was 32 Hz. At the same time, a series of shots of a bubbles cloud was taken by the camera. The obtained images were processed on a PC in the Bubble Wizard software. This software implements the photometric method for determining the size of the bubbles by the shadow analysis. The operation algorithm is described in work [8].

4 Results

The results were processed as follows. For each of the bands of the measured spectrum, the radius R_i, corresponding to the center frequency of the band f_i was calculated from the inverse Minnaert relation. Then, for each R_i, using expression (Eq. 20), we calculated the value of the distribution density function (R_i).

Photos of the bubble cloud were processed in the Bubble Wizard software. In this software, the result of the analysis is the Gaussian function as close as possible to the real distribution of bubbles.

Graphs of probability density functions calculated in two ways are presented in Fig. 3.

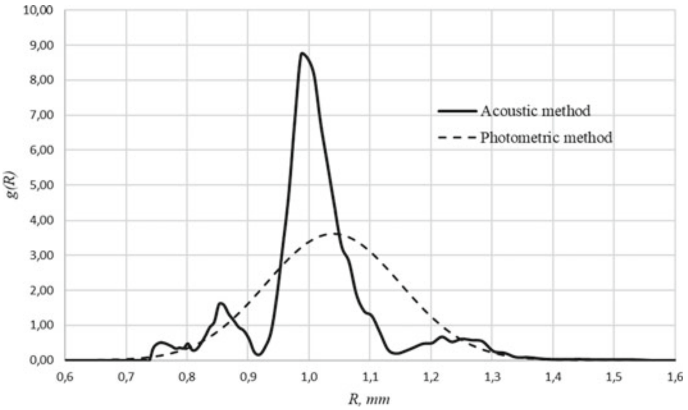


Fig. 3 Experiment Results. Distribution density functions obtained by acoustic and photometric methods

By the acoustic method, we obtained a close to normal distribution, but with two “humps” on the sides. If we compare this distribution with the results of the photometric method, then they have almost identical coverage of the radii. The most probable bubble size calculated by the acoustic method is 0.99 mm and photometric 1.04 mm. Since the relative error was less than 5%, it can be concluded that the proposed algorithm for converting the noise spectrum into the size distribution of bubbles works.

5 Conclusion

Currently, much attention is paid to the dispersed composition of the gas phase in a liquid medium, since this is a key parameter in the design and calculation of most hydromechanical systems. Passive acoustic methods are preferred for determining this parameter. They are based on the fact that air bubbles in water are sources of acoustic signals.

The work proposed a passive acoustic method for determining the distribution of air bubbles in water. The essence of the method is to measure the spectrum of bubble noise and convert it to a size distribution. Most of the existing passive methods allow determining only the sizes of single bubbles or bubbles, the size distribution of which has one “peak”. The mathematical model of the proposed method does not depend on the number and absence/presence of distribution peaks.

The correctness of the model was verified by comparison with the photometric method in an experimental setup. The comparison showed excellent convergence of the results obtained by acoustic and photometric methods. In further works, it is planned to study the parameters that affect the accuracy of the proposed method.

References

1. M.V. Ivanov, B.S. Ksenofontov, Intensification of flotation treatment by exposure to vibration. Water Sci. Technol. **69**(7), 1434 (2014)
2. A.A. Aleksandrov, V.A. Akat'ev, M.P. Tyurin, E.S. Borodina, O.S. Kochetov, Results of experimental studies of heat-and-mass transfer processes in a two-phase closed thermosyphon. Herald Bauman Moscow State Tech. Univ. Ser. Nat. Sci. **4**, 46 (2018)
3. B.S. Ksenofontov, V.P. Yakushkin, in *29th International Mineral Processing Congress IMPC 2018* (Canadian Institute of Mining, Metallurgy and Petroleum, Department of Ecology and Environment Protection, Bauman Moscow State Technical University, 2-Ya Baumanskaya, 5, Moscow, 105005, Russian Federation, 2019), pp. 1063–1074
4. B.I. Bakhtin, A.I. Ivashov, A.V. Kuznetsov, A.S. Skorokhodov, Experimental investigation of the specific features of formation of cavitation zones in intense ultrasound fields. J. Eng. Phys. Thermophys. **87**(3), 672 (2014)
5. B.V. Kichatov, V.M. Polyaev, S.V. Runkovskii, Gas bubble dynamics with instability in the operation of a system of transpiration cooling. Heat Transf. Res. **32**(4–6), 273 (2001)
6. Y. Zhang, H. Sun, Measurement of bubble size distribution in liquids by optical and acoustical methods, in *Proceedings—International Conference on Communication Systems and Network Technologies (CSNT 2012)* (2012), p. 671
7. M. Villiger, C. Pache, N. Bocchio, J. Goulley, T. Lasser, Biomed. Opt. (2010)
8. M.V. Ivanov, S.A. Gavrilev, J.M. Tyurina, A. Yusipova, M.D. Boldyrev, in *Lecture Notes in Engineering and Computer Science Proceedings of the World Congress on Engineering 2019*, London, UK, 3–5 July, 2019 (2019), pp. 299–303
9. T. Gaillard, C. Honorez, M. Jumeau, F. Elias, W. Drenckhan, A simple technique for the automation of bubble size measurements. Colloids Surf. A Physicochem. Eng. Aspects **473**, 68 (2015)
10. J.R. Hernandez-Aguilar, R.G. Coleman, C.O. Gomez, J.A. Finch, A comparison between capillary and imaging techniques for sizing bubbles in flotation systems. Miner. Eng. (2004)
11. X.J. Wu, G.L. Chahine, Development of an acoustic instrument for bubble size distribution measurement. J. Hydrodyn. **22**(5 SUPPL. 1), 325 (2010)
12. J. Xue, Bubble velocity, size and interfacial area measurements in bubble columns, No. December, 1 (2004)
13. M. Christensen, P. Thomassen, Experimental determination of bubble size distribution in a water column by interferometric particle imaging and telecentric direct image method, No. June, 121 (2014)
14. W.A. Al-Masry, E.M. Ali, Y.M. Aqeel, Determination of bubble characteristics in bubble columns using statistical analysis of acoustic sound measurements. Chem. Eng. Res. Des. **83**(10), 1196 (2005)
15. T.G. Leighton, A.J. Walton, An experimental study of the sound emitted from gas bubbles in a liquid. Eur. J. Phys. **8**(2), 98 (1987)
16. T.G. Leighton, R.J. Lingard, A.J. Walton, J.E. Field, Acoustic bubble sizing by combination of subharmonic emissions with imaging frequency. Ultrasonics **29**, 319 (1991)
17. M. Strasberg, Gas bubbles as sources of sound in liquids. J. Acoust. Soc. Am. **28**(1), 20 (1956)
18. K. Vokurka, On Rayleigh's model of a freely oscillating bubble. I. Basic relations. Czechoslov. J. Phys. **35**(1), 28 (1985)
19. T. Leighton, Acoustic bubble detection—I: the detection of stable gas bodies (1994)
20. J.W.R. Boyd, J. Varley, The uses of passive measurement of acoustic emissions from chemical engineering processes. Chem. Eng. Sci. **56**(5), 1749 (2001)
21. R. Manasseh, R.F. LaFontaine, J. Davy, I. Shepherd, Y.-G. Zhu, Passive acoustic bubble sizing in sparged systems. Exp. Fluids **30**(6), 672 (2001)
22. C. Greene, P.S. Wilson, R.B. Coffin, Laboratory measurements on gas hydrates and bubbly liquids using active and passive low-frequency acoustic techniques, No. May, 045001 (2011)
23. C.A. Greene, P.S. Wilson, Laboratory investigation of a passive acoustic method for measurement of underwater gas seep ebullition. J. Acoust. Soc. Am. **131**(1), EL61 (2012)

24. S. Husin, D. Mba, Correlation between Acoustic Emission (AE) and Bubble Dynamics. **II** (2010)
25. S.A. Gavrilev, M.V. Ivanov, E.A. Yusupov, Measurement of bubble size distribution by passive acoustic method, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, London, UK, 3–5 July, 2019, pp. 527–530

Movement Detection and Moving Object Distinction Based on Optical Flow for a Surveillance System

Paulo A. S. Mendes and A. Paulo Coimbra

Abstract Detection of moving objects in sequences of images is an important research field, with applications for surveillance, tracking and object recognition among others. An algorithm to estimate motion in video image sequences, with moving object distinction and differentiation, is proposed. The motion estimation is based in three consecutive RGB image frames, which are converted to gray scale and filtered, before being used to calculate optical flow, applying Gunnar Farnebäck's method. The areas of higher optical flow are maintained and the areas of lower optical flow are discarded using Otsu's adaptive threshold method. To distinguish between different moving objects, a border following method was applied to calculate each object's contour. The method was successful detecting and distinguishing moving objects in different types of image datasets, including datasets obtained from moving cameras. This extended version contemplates more results obtained, using the demonstrated methodology, with other datasets.

Keywords Computer vision · Image processing · Movement detection · Object detection · Optical flow · Robot vision

1 Introduction

Nowadays, more than ever, it is possible the development and application of more complex algorithms using inexpensive hardware. Optical flow and image processing in general are computationally demanding. Nonetheless, they are already possible in real time using common hardware.

P. A. S. Mendes (✉)
Institute of Systems and Robotics (ISR), Coimbra, Portugal
e-mail: 33paulomendes@gmail.com
URL: https://isr.uc.pt/

A. Paulo Coimbra
Department of Electrical and Computer Engineering,
University of Coimbra and ISR, Coimbra, Portugal
e-mail: acoimbra@deec.uc.pt
URL: https://www.uc.pt/en/fctuc/deec

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_12

Optical flow is an important method for motion estimation in visual scenes. Lucas and Kanade image registration method, also known as gradient-based optical flow, makes motion estimation in images possible with low computation time [1, 2]. Pyramidal Lucas and Kanade [3], Farnebäck [4–6] and Brox et al. [7] optical flow are other methods for motion estimation in images.

Sengar and Mukhopadhyay developed excellent methods to detect movement [8] and a moving object area [9], based on optical flow. The methods show precise results with low processing time, which is important for automatic surveillance and the detection of moving objects using computer vision. The method creates a smaller image, which is a fraction of the original image, and contains a representation of the moving objects detected. When there are multiple moving objects the image returned contains all the objects and requires further processing to distinguish between the objects.

Chen and Lu proposed object-level motion detection from a moving camera [9], estimating the objects' movement relative to the camera.

The present paper describes a method of motion estimation based on Gunnar Farnebäck's optical flow. The aim is to calculate a dense optical flow from consecutive images to accomplish the most precise movement detection in the least amount of time. Images are preprocessed, namely converted to gray scale and smoothed to improve the results. Otsu's threshold method [10] is applied to create a binary image representing the areas of larger optical flow. Suzuki and Abe's border following method [11] is applied to the binary image to make the distinction of different moving objects.

The paper presents results of different experiments using images selected from popular datasets, each one representing particular challenges for optical flow calculation. It is an extended version of the previous work developed in "Movement Detection and Moving Object Distinction Based in Optical Flow" [12] based in "Movement, Pedestrian and Face Detection Based on Optical Flow for Surveillance Robot" master thesis [13].

Section 2 describes the hardware, the software and the datasets used. The main steps of the algorithm are described in Sect. 3. The tests and results are presented and discussed in Sects. 4 and 5. Section 6 draws some conclusions and possible future developments.

2 Experimental Setup

2.1 Hardware

The hardware used was a laptop with a 2.40 GHz Intel Core i7-3610QM processor, 6 GB RAM and a NVIDIA Graphics Processing Unit (GPU) with 2 GB memory and 96 CUDA cores.

2.2 Datasets

The proposed methodology were tested using images from datasets adequate for optical flow calculation. The first experiments were performed using images from dataset "O_SM_02" from LASIESTA Database, of the Universidad Politécnica de Madrid [14], which contains outdoor images of moving people taken with a moving camera. The camera resolution is 352 × 288 pixels. Other datasets used are LASIESTA Database "O_SM_07", Freiburg University "Chinese Monkey" [15], INRIA CAVIAR[1] and Middlebury University "Evergreen" and "Dumptruck" [16].

2.3 Software

The algorithms were coded in C and C++ language using OpenCV (Open Source Computer Vision Library) [17]. OpenCV, originally developed by Intel, is a software toolkit for processing real-time image and video, that also provides analytics and machine learning capability. It is free for academic and commercial use. For a better result, part of the algorithm, was developed using OpenCV CUDA module. OpenCV CUDA module is a set of classes and functions to utilize CUDA computational capabilities.

Figure 1 shows a block diagram of the system architecture, as well as the interactions between camera or Hard Disk Drive (HDD) and the image processing module. The software has been implemented in a modular way. The images can be captured from a camera in real time, or read from the computer disk. The experiments described below use images from datasets, stored in the computer hard disk. The images are then sent to the image processing module. This module returns a list of moving objects, which can then be used by an "action selector". This algorithm was developed to make future applications of moving object detection possible in real time.

The algorithm to detect moving objects is represented in the flow chart shown in Fig. 2. This will be explained in Sect. 3.

3 Moving Objects Detection

Motion estimation is based on three consecutive RGB frames (F_1, F_2 and F_3), as shown in Fig. 3. These images will be preprocessed and then used for optical flow calculation as presented next.

[1] INRIA CAVIAR database: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/ (last checked 2020-09-27).

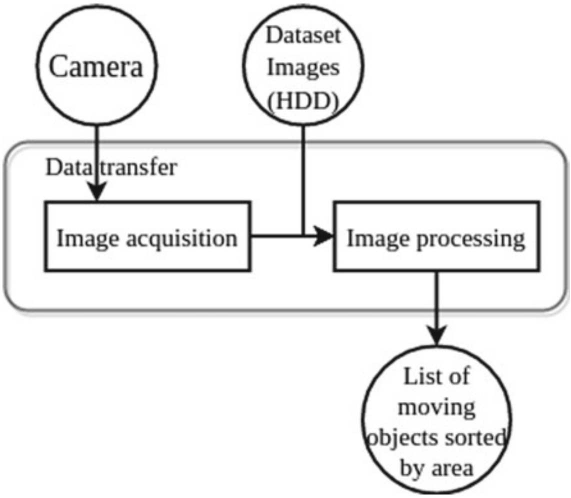


Fig. 1 System architecture with input images from a camera or an HDD and the system output (list of moving objects sorted by area)

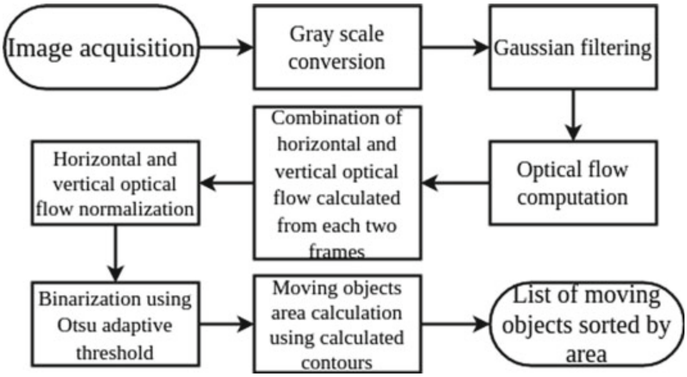
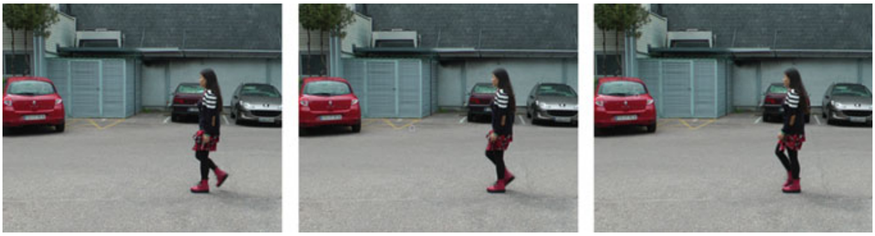


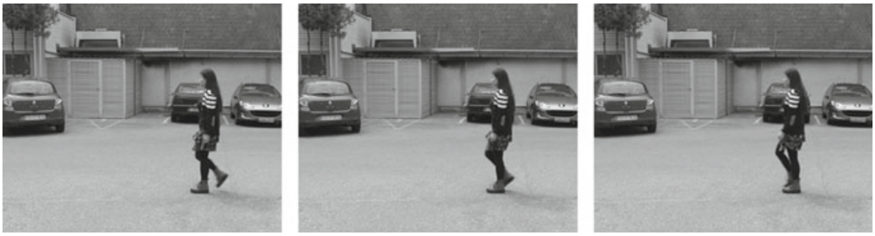
Image acquisition
Gray scale conversion
Gaussian filtering
Optical flow computation
Combination of horizontal and vertical optical flow calculated from each two frames
Horizontal and vertical optical flow normalization
Binarization using Otsu adaptive threshold
Moving objects area calculation using calculated contours
List of moving objects sorted by area

Fig. 2 Steps of the algorithm to detect moving objects



(a) Frame F_1 (b) Frame F_2 (c) Frame F_3

Fig. 3 Three consecutive RGB frames from LASIESTA database



(a)Frame $Gray_1$ (b)Frame $Gray_2$ (c)Frame $Gray_3$

Fig. 4 Gray scale images obtained from the RGB frames shown in Fig. 3

3.1 Conversion to Gray Scale

The three RGB frames are first converted to gray scale images. The conversion is made using Eq. 1, giving different weights to each color channel. The weights are the default OpenCV values.

$$Gray_i(x, y) = 0.299R_i(x, y) + 0.587G_i(x, y) + 0.114B_i(x, y) \tag{1}$$

R_i, G_i and B_i are the red, green and blue channel components of RGB frame F_i. The pixel position in the frame image is defined by coordinates (x, y). Figure 4 shows the same images as Fig. 3 after gray scale conversion, using Eq. 1.

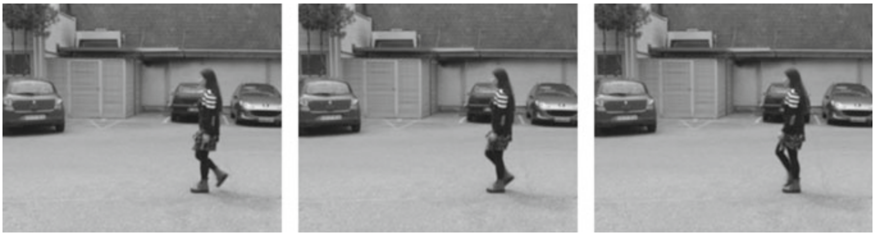
3.2 Noise Smoothing

Digital images usually have some noise. To minimize that problem, a Gaussian filter is applied to each frame using the two dimensional Gaussian function shown in Eq. 2.

$$Gaussian(x, y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \tag{2}$$

The resulting Gaussian distribution resembles a bell curve which is used to smooth the image. The standard deviation of the Gaussian filter distribution, σ, can be interpreted as a measure of its size, controlling the bell curve aperture. The Gaussian distribution is approximated to a suitable convolution kernel (a matrix composed of floating point values). To obtain the smoothed image it is necessary to convolve the Gaussian filter with the gray image. The convolution follows Eq. 3, using the convolution kernel that results from a chosen σ value.

$$Smooth_i(x, y) = Gaussian(x, y) * Gray_i(x, y) \tag{3}$$



(a)Frame $Smooth_1$ (b)Frame $Smooth_2$ (c)Frame $Smooth_3$

Fig. 5 Gray scale frames filtered with Gaussian filter

The kernel size was chosen to be of 3 × 3 for fast computation applying the filter.

The optimal value for the standard deviation, σ, varies from image to image. However, its calculation takes time. In the present work was used a constant value $\sigma = 1.5$ that showed good results in the experiments. Figure 5 shows the results obtained after filtering the images shown in Fig. 4, using Eq. 3.

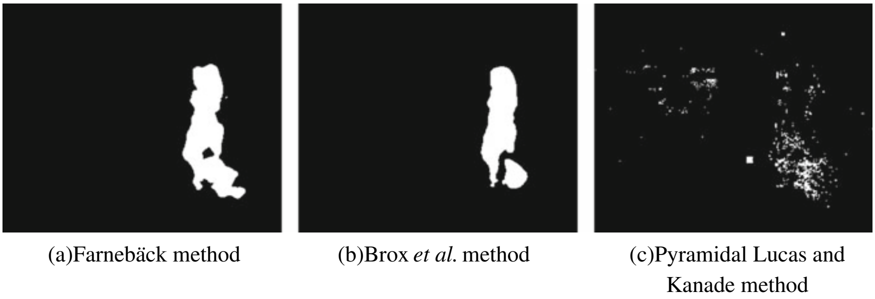
3.3 Optical Flow Computation

Optical flow output is a two dimensional (2D) field that represents moving objects in the real world or a moving camera taking frames of a scene. In computer vision, the main method for motion estimation is optical flow. It is also used in the present paper.

Pyramidal Lucas and Kanade method produces good results, at the cost of a significant amount of computation time. Brox et al. optical flow outperforms the other methods in precision. However, it also requires more processing time. Since the aim of the present work is to apply the developed algorithm in real time, Brox et al. optical flow was discarded. Farnebäck's method was also tested. It is faster than Pyramidal Lucas and Kanade. It showed the best results, in a precision to computation time ratio basis. Figure 6, shows the images of the optical flow obtained for these methods applied to the sequence of images of Fig. 5. Table 1 shows the computation times measured to calculate the optical flow among three consecutive images—that is to calculate the optical flow between frames F_1 and F_2 and between frames F_2 and F_3.

The aim of the present work is to detect movement in a sequence of images. Therefore, a dense optical flow calculation is applied for every two consecutive frames, using Gunnar Farnebäck's method.

The optical flow calculation results in horizontal and vertical directions optical flow images as shown in Fig. 7. Figure 7 show vertical and horizontal optical flows calculated using, respectively, the first and second frames and second and third frames shown in Fig. 5. The optical flow pixels are a projection of the motion field onto the $2D$ image. The motion field is a representation of the real $3D$ world motion.


(a)Farnebäck method

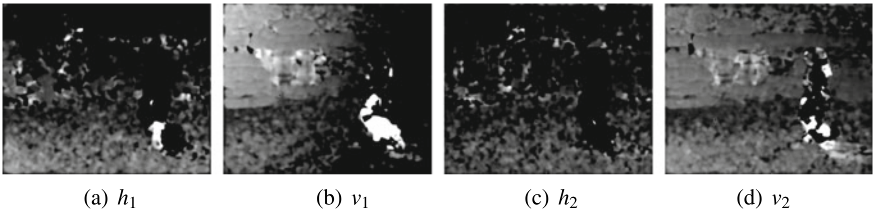
(b)Brox *et al.* method

(c)Pyramidal Lucas and Kanade method

Fig. 6 Resulting movement detection using different optical flow methods

Table 1 Computation times, in seconds, to calculate the optical flow between 3 consecutive frames using the CUDA GPU and different optical flow methods, for images of different datasets

Image set	Farnebäck	Brox *et al.*	Pyramidal Lucas and Kanade	Image size
"Dumptruck" set	0.154000	1.782000	0.320000	640 × 480
‘Evergreen" set	0.153700	1.756000	0.358000	640 × 480
INRIA "CAVIAR set 1	0.061000	0.770000	0.120000	384 × 288
INRIA "CAVIAR" set 2	0.057700	0.769000	0.127100	384 × 288
LASIESTA "O_SM_02" set	0.058870	0.736724	0.128260	352 × 288
LASIESTA "O_SM_07" set	0.060336	0.746460	0.115300	352 × 288
"Chinese Monkey" set	0.144000	1.887000	0.491800	720 × 432


(a) h_1

(b) v_1

(c) h_2

(d) v_2

Fig. 7 Optical flow in horizontal and vertical directions (h_1 and v_1), calculated between frames F_1 and F_2. Optical flow in horizontal and vertical directions (h_2 and v_2), calculated between frames F_2 and F_3

3.4 Optical Flow directions combination

For each two filtered consecutive frames there are the horizontal and vertical optical flow components, h_i and v_i, as explained above. Therefore, three consecutive frames (F_1, F_2 and F_3) result in four optical flow images (h_1, v_1, h_2 and v_2). The horizontal and vertical components are then added resulting into one single image for each optical flow direction, as shown in Eqs. 4 and 5.

$$H(x, y) = h_1(x, y) + h_2(x, y) \tag{4}$$

$$V(x, y) = v_1(x, y) + v_2(x, y) \tag{5}$$

H and V are the images of horizontal and vertical optical flow and (x, y) are the pixel coordinates. Figure 8 shows the results of applying Eqs. 4 and 5 to the images shown in Fig. 7a–d.

The magnitude of the horizontal and vertical optical flows is calculated applying Eq. 6.

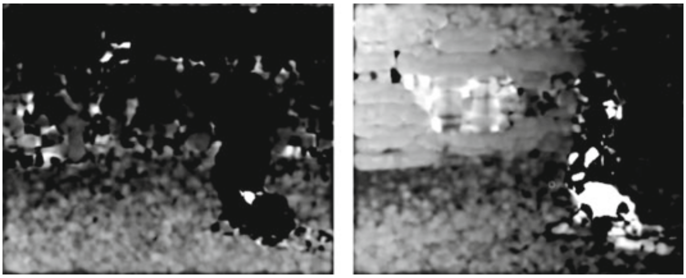
$$M(x, y) = \sqrt{H^2(x, y) + V^2(x, y)} \tag{6}$$

The resulting optical flow, obtained after applying Eq. 6 to images H and V, consists of one single 2D gray scale image (M), as shown in Fig. 9a.

The pixel's values in M do not occupy all the range of the gray scale images ([0, 255] for 8-bit gray scale images). Therefore, for a better result, the image M is normalized using Eq. 7.

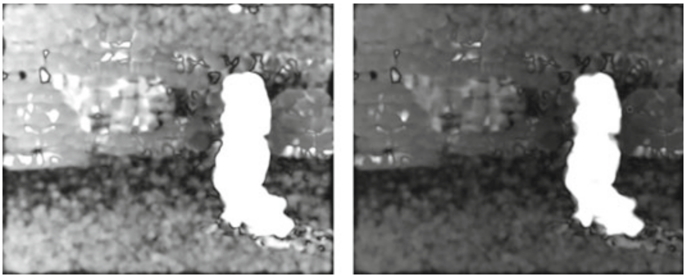
$$N(x, y) = \frac{M(x, y) - M^{min}}{M^{max} - M^{min}} \times I_{max} \tag{7}$$

In the equation, $M(x, y)$ is the optical flow magnitude value at pixel position (x, y). M^{min} and M^{max} are the minimum and maximum values in the image M. Finally,



(a)Horizontal optical flow (*H*) (b) Vertical optical flow (*V*)

Fig. 8 Sum of horizontal and vertical optical flows applying Eqs. 4 and 5 to the images shown in Fig. 7a–d



(a)Optical flow magnitude (M)

(b)Normalized optical flow magnitude (N)

Fig. 9 Optical flow magnitude following Eq. 6 and normalized optical flow following Eq. 7

$N(x, y)$, is the normalized optical flow at pixel position (x, y) and I_{max} is the maximum possible pixel value (255 for 8-bit gray scale images). Normalization facilitates the motion estimation procedure because of the wider range to calculate the optimal threshold value, as explained in next Section.

3.5 Movement Detection using adaptive threshold

The normalized optical flow image (N) contains noise (detected "false movement"). Noise which was not completely eliminated by the preprocessing of the original RGB frames (F_1, F_2 and F_3) causes "false movement" detection (noise) after optical flow calculation and normalization. This problem can be minimized by applying a threshold based segmentation method, as proposed by N. Otsu [10]. This method consists in computing the gray scale image histogram, and determining the optimal threshold value based on the histogram peaks. The threshold is chosen as a value between two peaks of a bimodal image histogram. The pixel values above the threshold value are set to one and the values below the threshold are set to zero.

Xu, Jin and Song refer that the threshold method is effective to separate objects from the background when the gray levels are substantially different between them [18]. Optical flow normalization, as described in Sect. 3.4, facilitates this task.

In the present work the threshold value, defined as λ, is calculated from the normalized optical flow image (N). Afterwards, the image is binarized applying Eq. 8.

$$B(x, y) = \begin{cases} 1 & \text{if } N(x, y) \geq \lambda \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The binarized image (B) resulting from applying Eq. 8 to image N is shown in Fig. 10. The white areas identify the areas of movement detected in the original sequence of three images. The black areas are areas where there is no movement detected.

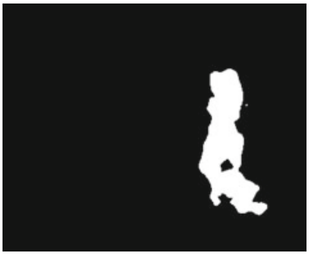


Fig. 10 Binarized image obtained applying Otsu's threshold method to the image of the Fig. 9b. The white area corresponds to the detected movement

3.6 Moving Object Area Detection

For a better distinction of the moving objects, the objects' contours are calculated from the binary image (B) and stored in a data structure (as a contour list). Figure 11 shows the two contours calculated from image B and drawn in a black background image. The contours are determined using Suzuki and Abe's border following method [11].

The list of contours contains small contours which are most likely noise. Therefore, the small contours are ignored. The area of all contours is calculated and only the contours which have an area larger than a predefined value are considered. Those contours represent some object (or part of one) moving in the real world. On the other hand, the contours with area smaller than the predefined value are too small to be relevant. In the present implementation, only the largest area contour is selected.

The area of the largest contour shown in Fig. 11 is 6342.5 $pixels$. It was calculated using Green's theorem[2] OpenCV implementation. Knowing the moving objects' contours, the corners of the minimum rectangle that contains those contours are calculated and the rectangle is marked in the original image, detecting the region of interest (ROI). Figure 12, shows a green rectangle over the ROI in the original RGB frames. The ROI cut out from each RGB frame can be seen in Fig. 13.

Figure 14 shows the moving person superimposed, cut out by hand, on the area of movement detected for comparative reasons. As the images show, the area where movement was detected contains approximately the union of the areas the person occupied in the three images. There is an extra area to the bottom right of the image which corresponds to part of the person's shadow, which also moved.

4 Results with Other Sets of Images

This section describes movement detection results obtained for images from different datasets. For each dataset the original three consecutive frames are shown, then the binarized image with the area of movement detected and the regions of interest (ROI).

[2] Green's theorem brief explanation: https://en.wikipedia.org/wiki/Green' stheorem (last checked 16.08.2018).

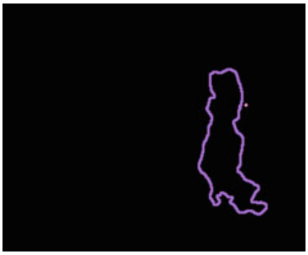
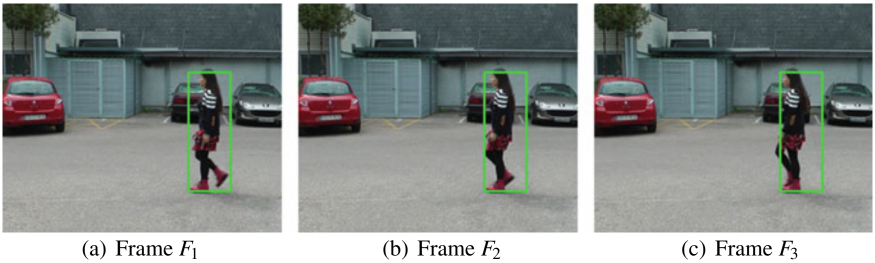


Fig. 11 Calculated contours from the binary image. The largest contour corresponds to a 6342.5 $pixels$ and the smaller to only 0.5 $pixels$ area



(a) Frame F_1 (b) Frame F_2 (c) Frame F_3

Fig. 12 Region of interest marked in green in the original RGB frames (F_1, F_2 and F_3) used to detect movement

Fig. 13 Cut out region of interest (ROI) from the largest area contour calculated, from each LASIESTA Database RGB frame (F_1,F_2 and F_3 frame)

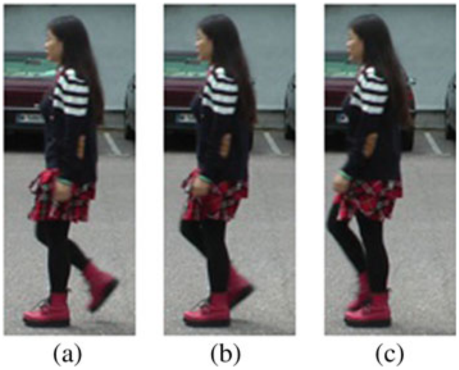


Figure 15 shows another set of frames from LASIESTA Database “O_SM_07”, which is an outdoor set with different types of camera motion. This set has medium camera jitter and low camera rotation. Figure 15d shows the binarized image presenting the detected movement areas. It is clear from this result that this set is difficult. Since the camera is non-static, not only the persons are moving but also the static objects in the scene. The threshold method eliminates the slow moving objects, but even so there are many areas with detected movement in the binary image. In these three frames sequence 97 contours were identified. These contours are shown in

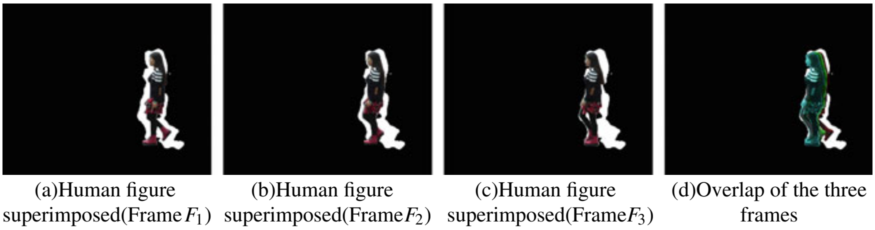

(a)Human figure superimposed(Frame F_1) (b)Human figure superimposed(FrameF_2) (c)Human figure superimposed(FrameF_3) (d)Overlap of the three frames

Fig. 14 Image of the human figure moving-object superimposed on the area of the detected movement, for each frame. Figure 14d show all three human figures (from each frame) superimposed

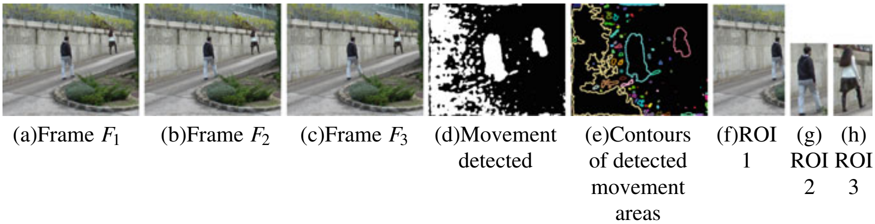

(a)Frame F_1 (b)Frame F_2 (c)Frame F_3 (d)Movement detected (e)Contours of detected movement areas (f)ROI 1 (g) ROI 2 (h) ROI 3

Fig. 15 LASIESTA database RGB image set ("O_SM_07"): **a**, **b**, **c** consecutive RGB frames used, **d** binarized image of movement detection, **e** contours of moving object areas and **f**, **g**, **h** ROI of the detected movement (from the three largest area contours)

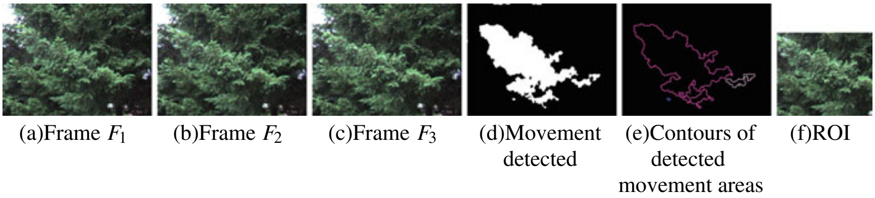
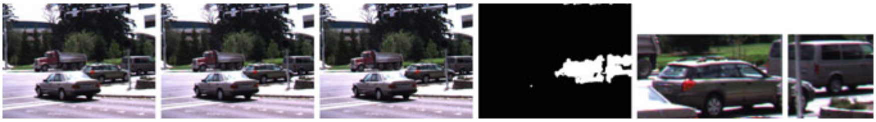

(a)Frame F_1 (b)Frame F_2 (c)Frame F_3 (d)Movement detected (e)Contours of detected movement areas (f)ROI

Fig. 16 Middlebury database RGB image set ("Evergreen"): **a**, **b**, **c** consecutive RGB frames used, **d** binarized image of the detected movement, **e** contours of the moving object and **f** ROI of the movement detection (from the largest area contour)

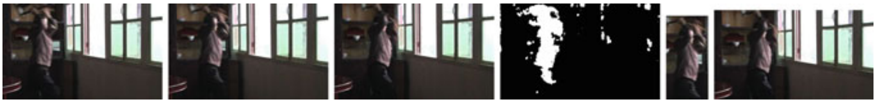
Fig. 15e. The areas of the three largest contours calculated using Green's theorem are 24325.5 (yellow), 5006.0 (cyan) and 2060.0 (pink) $pixels$. These contours results in the ROI seen in Fig. 15f–h respectively.

Figure 16 shows three RGB frames from Middlebury University Evergreen dataset. This set is a difficult one for movement detection because almost all images have large areas of the same color tone, namely green. In the three consecutive RGB frames the front branch moves with the wind, resulting in a movement detection of that branch, as shown in Fig. 16d. This image shows that despite the difficulty associated with the dataset, the method was successful in distinguishing the object with the largest movement from the other objects with negligible movements. Figure 16f shows the resulting ROI from the biggest contour calculated, seen in Fig. 16e.



(a)Frame F_1 (b)Frame F_2 (c)Frame F_3 (d)Movement detected (e)ROI1 (f)ROI2

Fig. 17 Middlebury database RGB image set ("Dumptruck"): **a**, **b**, **c** consecutive RGB frames used, **d** binarized movement detected and **e**, **f** ROI from the detected movement (from the two largest area contours)



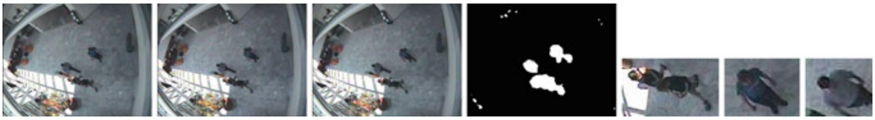
(a)Frame F_1 (b)Frame F_2 (c)Frame F_3 (d)Movement detected (e) ROI (f)ROI using Sengar *et al.* algorithm

Fig. 18 Freiburg database RGB image set ("Chinese monkey"): **a**, **b**, **c** consecutive RGB frames used, **d** binarized image of movement detected and **e** ROI from movement detection obtained with proposed methodology and **f** obtained using Sengar et al. algorithm

Figure 17 shows three consecutive RGB frames from Middlebury University, the Dumptruck dataset. The images were taken with a high-speed camera and show a street with four vehicles where two vehicles are moving fast while the other two are moving very slowly. Figure 17d shows the areas of movement detected. The largest area contour represents the Station Wagon vehicle moving and the ROI corresponding to that contour is shown in Fig. 17e. The second largest area contour represents the Van vehicle seen in Fig. 17f. Other contours have negligible area and are not considered.

Figure 18 shows three consecutive RGB frames from Freiburg university, the Chinese Monkey dataset. The Chinese Monkey dataset is a hard one because of the fast camera movement. The camera movement causes areas of motion in the image where there are only static objects. Since the algorithm applied to calculate the ROI was the contour of the largest area, the resulting ROI (from the movement detected seen in Fig. 18d) is only the area where there is a person, as shown in Fig. 18e. In this case the algorithm developed to differentiate regions of movement is more effective selecting the region of interest than the algorithm developed by Sengar and Mukhopadhyay as mentioned in Sect. 1. Using Sengar et al.'s method of movement detection, the result is almost all the original image frame, since there are regions of strong optical flow all over the analysed image (seen in Fig. 18d), as shown in Fig. 18f.

Figure 19 shows other three RGB frames from INRIA CAVIAR dataset. In those images there is a group of four people walking to meet at one location. The movement detection algorithm output can be seen in Fig. 19d. The ROI from the contours with a significant area are shown in Fig. 19e–g. In this case, the largest area contour represents the movement detected from the two people walking side by side.



(a)Frame F_1 (b)Frame F_2 (c)Frame F_3 (d)Movement detected (e)ROI1 (f)ROI 2 (g) ROI3

Fig. 19 INRIA CAVIAR database RGB image set ("Crowd of four people meet, walk and split"): **a**, **b**, **c** consecutive RGB frames used, **d** binarized image and **e**, **f**, **g** ROI from detected movement (from the three largest areas contours)

4.1 Computation Time

The methodology described in Sect. 3 was tested with the CPU and GPU capabilities, applied to the LASIESTA "O_SM_02" dataset. The total execution time measured, using the CPU, is of 114.218 milliseconds. The total execution time measured, using the GPU, is of 61.591 milliseconds. In Table 1 is shown the computation time for each one of the presented datasets, using the developed methodology with a CUDA GPU.

5 Discussion

Some surfaces, like metal grilles, are difficult for optical flow calculation. They result in false movement detection. In the present work the problem was minimized by applying a Gaussian filter as described in Sect. 3.2.

As shown in Sect. 3.3, Gunnar Farnebäck's optical flow calculation method shows good results within acceptable calculation time. That makes it appropriate for surveillance purposes where near real time processing is necessary. Otsu's method is a good choice for the calculation of the normalized optical flow threshold value. There are several improvements to that same method that could be applied in future work [18]. As described in Sect. 3.5 the threshold method is very important to reduce false movement detection (noise due to difficult surfaces on real world or a larger movement of the camera). The more correct the threshold value the more precise the movement detection will be. Suzuki and Abe's method to calculate contours in binary images showed good results in the present work. The contours facilitate the distinction between moving objects. This is specially important in frames taken with a moving camera.

Experimental results show the methodology proposed was effective detecting moving objects even when the camera is moving (pan, tilt, jitter and rotation).

6 Conclusion

A new methodology capable of detecting different moving objects using Gunnar Farnebäck's optical flow, Otu's adaptive threshold and Suzuki and Abe's contour calculation method is presented.

The methodology was successful in detecting distinct moving objects in images from static and moving cameras. Regions of the image can then be selected based on the moving objects' contour areas. Experiments using images from different datasets show that the methodology is effective and fast enough to be used in real time. Therefore, it is suitable to use in applications such as surveillance, object tracking, object counting and others applications.

Future work includes development of an optimal Gaussian filter calculator for gray scale images, in order to get a better optical flow precision. Otsu's threshold method can also be improved, as discussed in Sect. 5. The process can also be sped-up using parallel GPU computation. As mentioned in Sect. 5, it is possible to use another optical flow calculation method to achieve even better precision on the movement detection at the cost of additional calculation time.

References

1. B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, *International Joint Conference on Artificial Intelligence (7th)*, vol. 2, pp. 674–679 (1981)
2. A. Faria, Fluxo ptico, *ICEx-DCC-Visao Computacional* (1992)
3. J.-Y. Bouguet, Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 10 (2000)
4. G. Farnebäck, Two-frame motion estimation based on polynomial expansion, *Scandinavian Conference on Image Analysis*, pp. 363–370 (2003)
5. G. Farnebäck, *Polynomial expansion for orientation and motion estimation*, Ph.D. dissertation, Department of Electrical Engineering, Linköping University, Sweden (2002)
6. G. Farnebäck, Fast and accurate motion estimation using orientation tensors and parametric motion models, *International Conference on Pattern Recognition. ICPR-2000*, vol. 1, pp. 135–139 (2000)
7. T. Brox, A. Bruhn, N. Papenberg, J. Weickert, High accuracy optical flow estimation based on a theory for warping, *Lecture Notes in Computer Science*, vol. 3024, pp. 25–36 (2004)
8. S.S. Sengar, S. Mukhopadhyay, Detection of moving objects based on enhancement of optical flow. Optik **145**, 130–141 (2017)
9. S.S. Sengar, S. Mukhopadhyay, Moving object area detection using normalized self adaptive optical flow. Optik **127**(16), 6258–6267 (2016)
10. N. Otsu, A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
11. S. Suzuki, K. Abe, Topological structural analysis of digitized binary images by border following. Comput. Vis. Graph. Image Process. **30**(1), 32–46 (1985)
12. P.A.S. Mendes, M. Mendes, A.P. Coimbra, M.M. Crisóstomo, Movement detection and moving object distinction based on optical flow, *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, 3–5 July (London, UK, 2019), pp. 48–53

13. P. A. S. Mendes, *Movement, Pedestrian and Face Detection Based on Optical Flow for Surveillance Robot*. Master's thesis, Department of Electrical and Computer Engineering, Faculty of Sciences and Technology, University of Coimbra, Portugal, September 2018
14. T. Brox, A. Bruhn, N. Papenberg, J. Weickert, Labeled dataset for integral evaluation of moving object detection algorithms: Lasiesta. Comput. Vis. Image Underst. **152**, 103–117 (2016)
15. T. Brox, J. Malik, Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(3), 500–513 (2011)
16. S. Baker, D. Scharstein, J. Lewis, S. Roth, M.J. Black, R. Szeliski, A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
17. OpenCV, *Open Source Computer Vision Library*. https://opencv.org/. Last checked 16 Aug 2018
18. X. Xu, S. Xu, L. Jin, E. Song, Characteristic analysis of Otsu threshold and its applications. Pattern Recogn. Lett. **32**(7), 956–961 (2011)

Data Security with DNA Cryptography

Anupam Das, Shikhar Kumar Sarma, and Shrutimala Deka

Abstract In the present day world, a lot of works have been done in making the data communication safe and secured. But an illegal professional practice, i.e. stealing the data during communication is still going on and the efforts in this field are constantly made to intrude into the network to crack the encrypted data before reaching it to the authenticated destination by some black hat persons. On the other hand, there are lots of researches are going on for making the data secured by encrypting them during communication and an efficient way of generating key to decrypt the encrypted data. There are so many techniques for encryption-decryption, i.e. cryptographic methods are used to make the data safe and secured during transmission. Here we are analyzing a cryptographic technique which is used earlier by some eminent scholars. But in those works the input-output fragments, analysis of the nature of the output generated and the details of the findings from the entire mechanism were missing. Here we discussed it clearly so that it made easy for the future researchers in this field and now they can take this work further more. This is the primary motive of this work and in this paper, we worked very hard to explain the various aspects of the DNA cryptography and its working and also an honest attempt is made to provide the important modules which are used. The algorithm is implemented using C++ and finally some examples of input-output are given for final analysis and conclusion.

Keywords Codon · DNA · Encryption · Decryption · Cryptography · Cipher-text · Password · Message · Key

A. Das (✉)
Cotton University, Gauhati, Assam, India
e-mail: adas_arya@rediffmail.com

S. K. Sarma
Gauhati University, Gauhati, Assam, India
e-mail: sks@gauhati.ac.in

S. Deka
CDAC, Kolkata, West Bengal, India
e-mail: sruti.mail20@gmail.com

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_13

1 Introduction

THIS paper deals with Data security with DNA cryptography. DNA is the abbreviated name for deoxyribonucleic acid which is the store house of all the genetic information of living organisms. The information stored in the genes within the DNA are instructions that tell the body how to construct that organism cell by cell. DNA is shaped in a double helix consisting of two complementary strands that bond to form the final structure. The most basic building block of DNA are the four nitrogen bases, namely, Adenine, Guanine, Cytosine and Thymine. These can bond in a particular fashion and form unique sequences of protein strands. The complementary bases are Adenine and Thymine, and Guanine and Cytosine.

DNA cryptography is the latest technology in cryptographic methods where the natural process of DNA formation has been used to encrypt information and then retrieve them by decrypting it. The biological structure of DNA is such that once information has been transformed into the basic forms of the four nitrogen bases, the process of protein formation.

2 The Problem Definition

2.1 Discussion on the Problem

The cryptographic algorithms that already exist have the common strategy to have a large keyspace and a complicated algorithm. For symmetric cryptography, the use of one time pad is the most simple solution to the key distribution problem. However, with increasing advancement in technology, it is getting easier to break the algorithms that are widely in use. The increasing length of OTPs are also a cause of concern.

For a more secure data hiding and symmetric key generation using genetic database, DNA cryptography has been proposed.

The data security and protected communication among Mobile Node (MN) and Correspondent Node (CN). This algorithm detects and prevents an attacker who intends to modify the data by using a suitable existing encryption algorithm [1]. The research is also done to map CRC cards into stochastic petri net for evaluating and analyzing quality parameter of security [2]. In another way a method is implemented where a model of security, including control of user access to databases of big data with RMS, the multiplicity and the virtual machine to prevent internal threats, deleting data, insecure or incomplete data protection and control of a third-party can be provided to improve the operation according to the rules of Petri net modeling and simulation [3].

2.2 What Is DNA?

DNA is the abbrviation of deoxyribonucleic acid. It is a moluecule with a long structure which consists of the unique code called genetic code of any living being. As an insrtuction manual contains the steps and rules for any process the DNA holds the instructions of all the proteins of the bodies of any living beings. This unique code reserves all the characteristics of living beings. This DNA makes every individual unique and this uniqueness is carried in the DNA from the parents to the childs and so to the subsequent hierrachies. All individuals have their own DNA structure as no two individuals are equal, even twins are having unique DNA structures.

2.3 Some Distinguished Characteristics of DNA

(i) DNA is responsible to make GENOME
(ii) The four basic block of DNA are: Adenine (A), Cytosine (C), Guanine(G), and Thymine(T).
(iii) The GENOME gets instruction from the sequence of the basic bases of DNA.
(iv) A, C, G and T make the strand of the DNA
(v) Deoxyribonucleic acid is a two stranded molecule.
(vi) In DNA, the strands are “double helix” shaped and twisted within.
(vii) DNA molecule with its complementary bases form “rungs”.
(viii) The combining mechanism is always same as A combines with T and C combines with G.
(ix) The joining element of the base is hydrogen.
(x) Francis Crick and James Watson found the double helix structure of DNA with the help of the two DNA scientists Rosalind Frankline and Mourice Wilkins.
(xi) All living beings have different sizes of GENOME, human being’s GENOME size is 3.2 billions (Fig. 1).

2.4 Advantages of Computing Copyright DNA Structure

(i) Speed: The conventional computer can compute at the rate of 100 millions of instructions per second (MIPS) approximately but experimentally it is found that DNA strands combinations are generated by combining DNA strands on computing at the rate of 109 MIPS or 100 times faster than a fastest computer.
(ii) Storage: The media storage requires 10 12 cubic nanometer to store 1 bit but DNA needs only 1 bit per cubic nanometer.
(iii) Power requirements: Since the DNA computing is based on chemical bonds and structures it does not need any outside power.

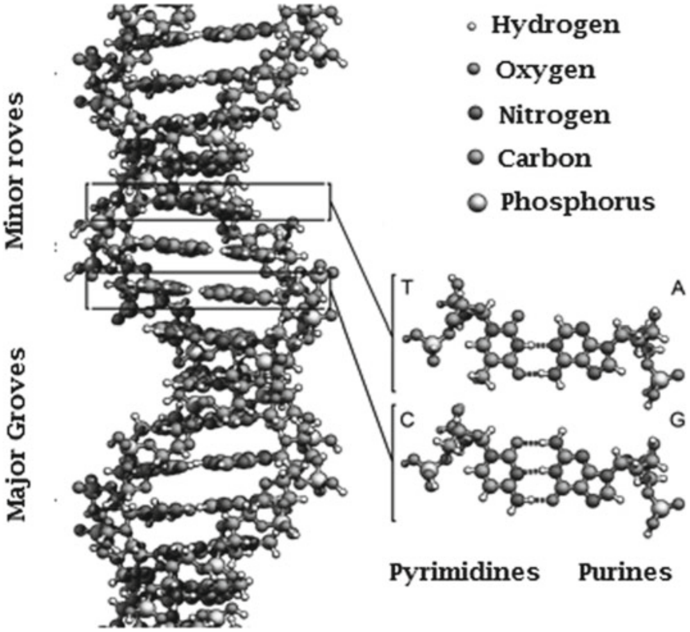


Fig. 1 DNA Structure *Source* https://en.wikipedia.org/wiki/File:DNA_Structure%2BKey%2BLabelled.pn_NoBB.png

2.5 Advantages of DNA Storage of Data

(i) Medium of Ultra-compact Information storage: Very large amounts of data that can be stored in compact volume.
(ii) A gram of DNA contains 1021 DNA bases = 108 Terabytes of data.
(ii) A few grams of DNA may hold all data stored in the world (Fig. 2).

3 Implementation

In Implementing the modules of the DNA cryptography, the C++ is used.

3.1 Key Generation

Start

1. Take input string password, lower case with no spaces
2. Store integer value of each character of password

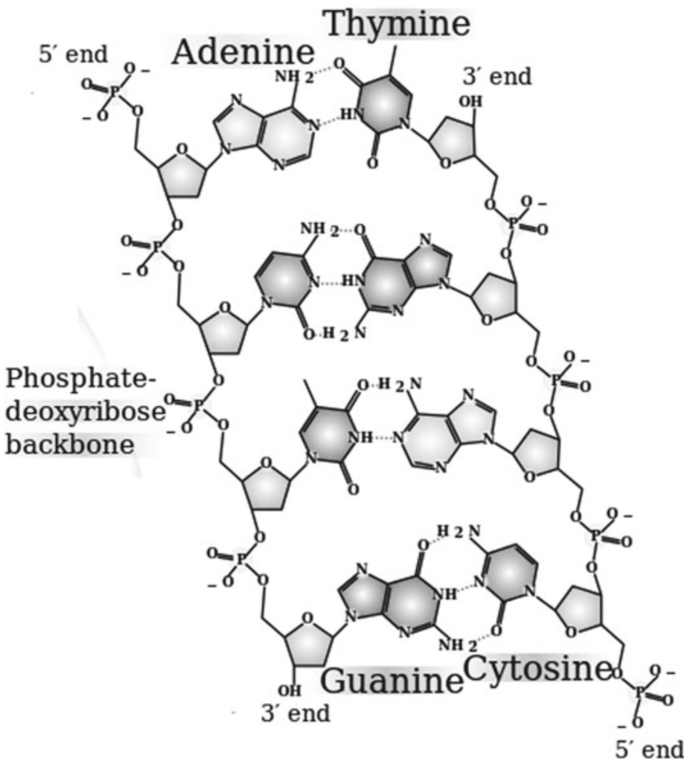


Fig. 2 DNA blocks or 'bases': Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) *Source* https://simple.wikipedia.org/wiki/DNA#/media/File:DNA_chemical_structure.svg

3. Convert to equivalent binary values (7 bits) and store in vector bitset structure named b_key
4. Take pair of binary bits in b_key from the right (LSB) and map them to nucleotides. Take the MSB as 0-bit.
5. Store in nucleotide vector
6. Perform annealing by concatenating the nucleotide string with another obtained by using complementary rule.
7. Perform transcription by mapping each T to U.
8. Parse the string for stop codons UAA, UGA, UAG and record their position.
9. Count the lengths of each string obtained between these stop codons starting from the beginning and ending at the last codon.
10. If multiple strings of various length obtained
    choose the longest
    Else if no codons are obtained choose the entire transcriped string
11. This is the protein key
12. Convert to binary bits and store into fkey
13. Output is fkey

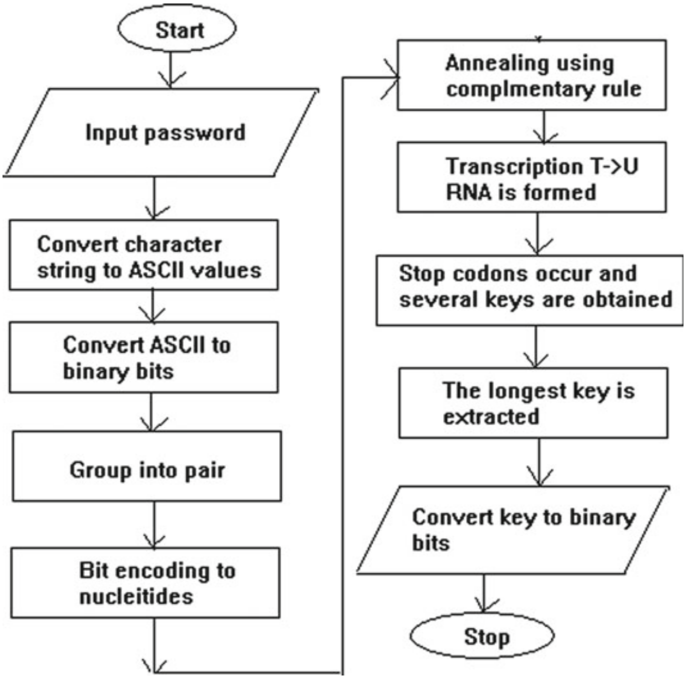

Start
Input password
Convert character string to ASCII values
Convert ASCII to binary bits
Group into pair
Bit encoding to nucleitides
Annealing using complmentary rule
Transcription T->U RNA is formed
Stop codons occur and several keys are obtained
The longest key is extracted
Convert key to binary bits
Stop

Fig. 3 Flowchart of key generation

Stop
[shown in Fig. 3].

3.2 Encryption

Start

1. Take input string Msg from the user
2. Store integer value of each character of Msg
3. Convert to equivalent binary values (7 bits) and store in vector bitset structure named b_msg
4. Perform circular left shift on each binary blocks in b_msg such that first block is shifted by 1 bit, second block by 2 bits and so on. Blocks which are multiple of 8 (8, 16..) are shifted starting from 1 bit again.
5. Xor b_msg with fkey and store in c_msg.
   If fkey is smaller than b_msg
   then repeat fkey blocks from the beginning
6. Output is encrypted message c_msg
   Stop

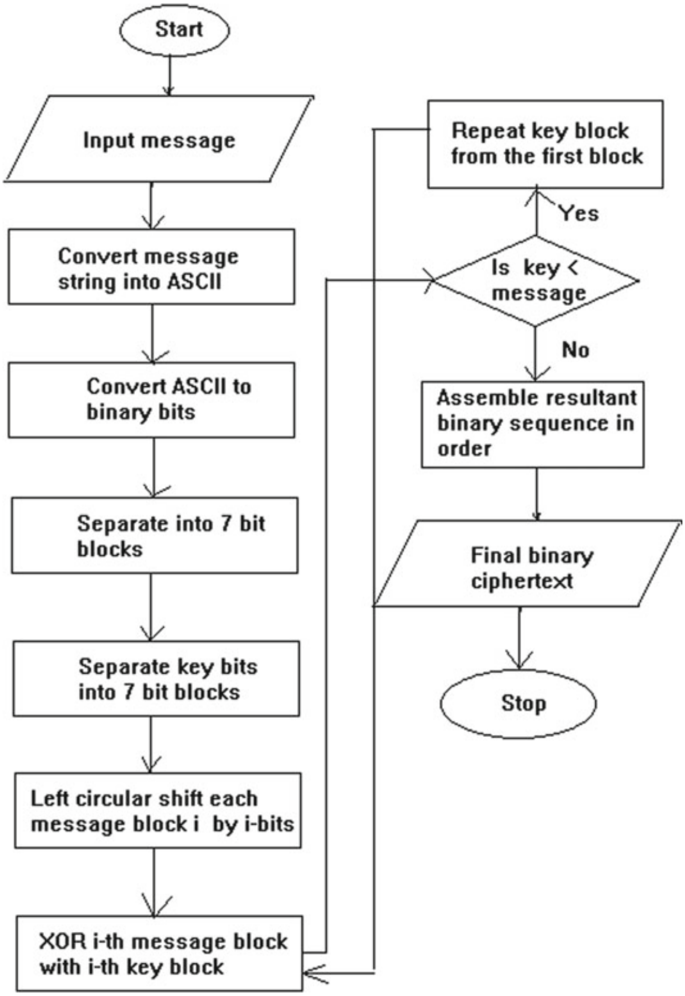

Start
Input message
Convert message string into ASCII
Convert ASCII to binary bits
Separate into 7 bit blocks
Separate key bits into 7 bit blocks
Left circular shift each message block i by i-bits
XOR i-th message block with i-th key block
Is key < message
Yes
No
Repeat key block from the first block
Assemble resultant binary sequence in order
Final binary ciphertext
Stop

Fig. 4 Flow chart of message encryption

[shown in Fig. 4].

3.3 Bit Encoding to Nucleotide

0 0 is used for A 1 1 is used for C
0 1 is used for G 1 0 is used for T

3.4 Anneal Using Complementary Rule

A [0 0] its complement is C [1 1]
G[0 1] its complement is T [1 0]

3.5 STOP Codons

UAG UAA UGA

Key generation module

```
Character string to binary bit string conversion
cout << “Enter password [space] Message: “;
cin ≫ password;
int len=password.length();
char Mpass[len+1];
strcpy(Mpass, password.c_str()); //string to char
     for(int t=0; t<len; t++)
     {char letter=Mpass[t];
     bitset<7>b(letter); //to binary b_key.push_back(b); //store into b_key
}
```

Obtaining nucleotides

```
(Here, comp parses through the bits)
int c=0;
  for(int t=0; t<b_key.size(); t++) {
  int i=0; while(i<7) {
  if(i==6) {comp.clear();
  comp.push_back(b_key[t] [6]);
if(comp[c]==0) nucleotide.push_back(‘A’);
else if(comp[c]==1) nucleotide.push_back(‘T’);}
  else {comp.clear(); comp.push_back(b_key[t][i]);
  comp.push_back(b_key[t][i + 1]);
  if(comp[c]==0 && comp[c + 1] ==0)
  nucleotide.push_back(‘A’);
  else if(comp[c]==0 &&
comp[c+1] ==1) nucleotide.push_back(‘G’);
else if(comp[c]==1 &&comp[c+1]==0)
  nucleotide.push_back(‘T’);
  else if(comp[c]==1 && comp[c+1]==1)
  nucleotide.push_back(‘C’); } i=i+2; }}
```

Extracting the longest key terminated by $

```
   (Here, p_key stores all the candidate keys that ends with
$)
int count1=0; //counts to $ encountered to choose zth key
     for(int t=0; t<count; t++) {//when first key is the
longest if(z ==0) //first key lies before the first ${while(p_key[t + 1] ! = '$') {
   key.push_back(p_key[t]);
t ++; } key.push_back(p_key[t]); //for last char
   break; } else {if(p_key[t] == '$')count1 ++; if(count1 ==z) //key is between
(z-1)th $ and the zth $
{while(p_key[t + 1] ! = '$'){key.push_back(p_key[t + 1]);
t ++;} break;} }}
```

Encryption function

```
Circular left shift operation
int e = 0;
   for(int t = 0; t < b_msg.size(); t ++) {
   if(t%7 == 0) {e = 0;} int shift = e + 1; e ++;
b_msg[t] = b_msg[t] < < shift | b_msg[t] > > (7-shift);
   cout < < b_msg[t] < < " ";//print
```

4 Input-Output

Example 1

```
Input password: apassword Input message: a message
Output : Password in binary bits: 11000011110000011000011
11001111100111110111110111111100101100100
nucleotide:TAGTAACTTAGTCACTCACTCTCTCCGT
ACTATGT
Annealed:TAGTAACTTAGTCACTCACTCTCTCCGT
ACTATGTATCATTGAATCAGTGAGTGAGAGAGGC
CTGATACA
Transcription:UAGUAACUUAGUCACUCACUCUCU
CGUGACUAUGUAUCAUUGAAUCAGUGAGUGAG
GAGGCACUGAUACA
Total Stop codons/Number of protein keys: 8
Codon at 0 Codon at 3 Codon at 8
Codon at 27 Codon at 41 Codon at 49
Codon at 53 Codon at 65
$$ CU$
UCACUCACUCUCUCCG$ CUAUGUAUCAU$
```

```
AUCAG$ G$ GAGAGGCAC$
Key length 1 : 0 Key length 2 : 0 Key length 3 : 2
Key length 4 : 16 Key length 5 : 11 Key length 6 : 5
Key length 7 : 1 Key length 8 : 9
Longest Key is: key 4 with length: 16
KEY: UCACUCACUCUCUCCG
Final binary key: 1010101 1000011 1000001 1000011
1010101 1000011 1000001 1000011 1010101 1000011
1010101 1000011 1010101 1000011 1000011 1000111
Binary Message: 1100001 1101101 1100101 1110011
1110011 1100001 1100111 1100101
Circular left shifting messages…
1000011 0110111 0101110 0111110 1111100 1110000
1100111 1001011
Encrypted message:
0010110 1110100 1101111 1111101 0101001 0110011
0100110 0001000 1010101
In Decryption….
Message BEFORE XORing with key: 1000011 0110111
0101110 0111110 1111100 1110000 1100111 1001011
0000000
Reversing circular shift(retrieve binary Message):
1100001 1101101 1100101 1110011 1110011 1100001
1100111 1100101 0000000
Decrypted message: a m e s s a g e
```

Example 2

```
PASSWORD MESSAGE
pswd End of Conversation
Password in binary bits:
1110000111001111101111100100
nucleotide: AACTCACTCTCTATGT
Annealed:AACTCACTCTCTATGTTTGAGTGAGAG
TACA
Transcription:AACUCACUCUCUAUGUUUGAGUG
GAGAUACA
Total Stop codons/Number of protein keys: 2
Codon at 17 Codon at 21
AACUCACUCUCUAUGUU$
G$ Key length 1 : 17 Key length 2 : 1
Longest Key is: key 1 with length: 17
KEY: AACUCACUCUCUAUGUU
Final binary key: 1000001 1000001 1000011 1010101
1000011 1000001 1000011 1010101 1000011 1010101
```

```
1000011 1010101 1000001 1010101 1000111 1010101
1010101
Binary Message: 1000101 1101110 1100100 1101111
1100110 1000011 1101111 1101110 1110110 1100101
1110010 1110011 1100001 1110100 1101001 1101111
1101110
Circular left shifting messages…
0001011 0111011 0100110 1111101 1011001 1100001
1101111 1011101 1011011 0101110 0101110 1111100
1110000 1110100 1010011 0111111 1110110
Encrypted message:
1001010 1111010 1100101 0101000 0011010 0100000
0101100 0001000 0011000 1111011 1101101 0101001
0110001 0100001 0010100 1101010 0100011
In Decryption….
Message BEFORE XORing with key: 0001011 0111011
0100110 1111101 1011001 1100001 1101111 1011101
1011011 0101110 0101110 1111100 1110000 1110100
1010011 0111111 1110110
Reversing circular shift(retrieve binary Message): 1000101
1101110 1100100 1101111 1100110 1000011 1101111
1101110 1110110 1100101 1110010 1110011 1100001
1110100 1101001 1101111 1101110
Decrypted message: E n d o f C o n v e r s a t i o n Example-
```

Example 3

```
PASSWORD MESSAGE
shortpass Meet on the Eastside
Password in binary bits:
111001111010001101111111001011101001110000110000
11100111110011
nucleotide:CACTAGGTCCGTGACTATCTAACTTAGT
ACTCACT
Annealed:CACTAGGTCCGTGACTATCTAACTTAGT
ACTCACTGTGATCCAGGCACTGATAGATTGAATC
GTGAGTGA
Transcription:CACUAGGUCCGUGACUAUCUAACU
AGUCACUCACUGUGAUCCAGGCACUGAUAGAU
GAAUCAGUGAGUGA
Total Stop codons/Number of protein keys: 10
Codon at 3 Codon at 11 Codon at 19 Codon at 24
Codon at 37 Codon at 49 Codon at 52 Codon at 57
Codon at 65 Codon at 69
CAC$ GUCCG$ CUAUC$ CU$ UCACUCACUG$
```

```
UCCAGGCAC$ $ AU$ AUCG$ G$
Key length 1 : 3 Key length 2 : 5 Key length 3 : 5
Key length 4 : 2 Key length 5 : 10 Key length 6 : 9
Key length 7 : 0 Key length 8 : 2 Key length 9 : 5
Key length 10 : 1
Longest Key is: key 5 with length: 10
KEY: UCACUCACUG
Final binary key: 1010101 1000011 1000001 1000011
1010101 1000011 1000001 1000011 1010101 1000111
Binary Message: 1001101 1100101 1100101 1110100
1101111 1101110 1110100 1101000 1100101 1000101
1100001 1110011 1110100 1110011 1101001 1100100
1100101
Circular left shifting messages…
0011011 0010111 0101110 1001110 1111011 0110111
1110100 1010001 0010111 0101100 0011100 1111100
0111010 1110011 1010011 0010011 0101110
Encrypted message:
1001110 1010100 1101111 0001101 0101110 1110100
0110101 0010010 1000010 1101011 1001001 0111111
1111011 0110000 0000110 1010000 1101111
In Decryption….
Message BEFORE XORing with key: 0011011 0010111
0101110 1001110 1111011 0110111 1110100 1010001
0010111 0101100 0011100 1111100 0111010 1110011
1010011 0010011 0101110
Reversing circular shift(retrieve binary Message):
1001101 1100101 1100101 1110100 1101111 1101110
1110100 1101000 1100101 1000101 1100001 1110011
1110100 1110011 1101001 1100100 1100101
Decrypted message: M e e t o n t h e E a s t s i d e
```

5 Theoretical Analysis

5.1 *Biological Aspect*

Placing an encrypted protein sequence among the vast number of protein strands poses the issue where the desired strand cannot be randomly located in the DNA. The key, thus, must also include where exactly is the encrypted message kept otherwise searching for the message itself will take too many years. Hence, adversaries with no knowledge of the key cannot possibly break the algorithm. Most strands only differ by few nucleotides. Without the key, it is impossible to even guess the cipher text, let alone decrypt it. This is a unique property of DNA cryptography and no modern cryptographic algorithm provides this kind of security in data.

5.2 Mathematical Aspect

Mathematical computations are minimal in DNA cryptography. This is because the role of confusion and diffusion are negligible, since the cipher text will not give away any clues to the plaintext. A large key space has been our solution so far to reduce the possibility of breaking cryptographic algorithms. This is eliminated in DNA cryptography as key is well hidden within the DNA. In other cases, we could also use genetic database to generate OTPs and eliminate the need to input password for the key altogether and keep the key space small at the same time.

5.3 Observations

(i) A single password can produce multiple protein keys in relation to the number of stop codons formed. On the other hand, a single password may also derive a single protein key. Then, we must use the one and only string as our key.

(ii) A given password will produce the same number of codons with the same number of keys and key length. This means that there is consistency in our output and it is not randomly generated each time. This also implies that using the same password again and again may not be such a good idea as the same protein key formed will become common knowledge with the people it is being shared with.

(iii) When the number of keys with the same maximal length is more than one, the program chooses the first longest key. This is done for convenience and has no particular reason in the security of the algorithm. If we decide to randomly choose to pick any one of the keys it will make the encryption even stronger with no scope of guessing the key.

(iv) The length of the password entered by the user is directly proportional to the number of stop codons or protein keys that we find. Heuristically it has been observed that for shorter passwords the number of key decreases. When the password is longer, the number of protein keys also increases. This does not mean that a password with length three will always produce less keys than that with length say five. On an average, the password length and number of keys are directly proportional.

(v) When the password is exceptionally small, as we have tested for the sake of proper output analysis, we may find that the stop codons produced is zero. This means that we may not even have a key. In that case, we take the entire annealed string as our key.

(vi) It has been found that usually the first or one of the first three proteins keys are found to be the longest. It is extremely rare that the last protein key be the longest.

(vii) In message encryption using the final key, the codons play no role in how the final encrypted string will look like. This is because the codons are part of only the key generation process and does not influence the rest of the processes.

(viii) Stop codons are UAA, UAG and UGA. Thus, for every occurrence of Thymine T, it becomes more likely that a stop codon will form in that position since every T will be transcriped into U in the subsequent steps in key generation function.

6 Conclusion and Future Work

The DNA Cryptography can now be used as the strong algorithm for data security as its cracking time and key generation are so designed that it seems the time taken to decrypt the ciphered data is quite impossible for a life time. So it should be the first choice for the cyber security researchers for securing data and information. The study made here is comprehensive and the information given here will largely help to the researchers for doing further work in this line of thinking. The modules given for key-generation, encryption, decryption will definitely help the subsequent works for implementing cryptographic techniques. The present work will also help to implement and apply DNA methodologies to cryptography and steganography.

References

1. A. Mehdizadeh, M. Mohammadpoor, Z. Soltanian, "Secured Route Optimization and Micromobility with Enhanced Handover Scheme in Mobile IPv6 Networks", in International Journal of Engineering (IJE). TRANSACTIONS B: Applications **29**(11), 1530–1538 (2016)
2. . H. Motameni a, M. Nemati b, Mapping, "CRC Card into Stochastic Petri Net for Analyzing and Evaluating Quality Parameter of Security", in IJE TRANSACTIONS B: Applications Vol. 27, No. 5, (May 2014) pp. 689–698
3. A.S. Abad, H. Hamidi, "An Architecture for Security and Protection of Big Data", in International Journal of Engineering (IJE). TRANSACTIONS A: Basics **30**(10), 1479–1486 (2017)
4. . B. B. Raj, J. Frank, T.Mahalakshmi, "Secure Data Transfer through DNA Cryptography using Symmetric Algorithm", in International Journal of Computer Applications, Vol 133-No 2, pp. 0975–8887, January 2016
5. A. Roy, A. Nath, "DNA Encryption Algorithms: Scope and Challenges in Symmetric Key Cryptography", in International Journal of Innovative Research in Advanced Engineering, ISSN: 2349–2763, Issue 11, Volume 3, Nov, 2016
6. W. Stallings, *Cryptography and Network Security* (Third Editio, Prentice Hall International, 2003)
7. N. S. Kolte, K. V. Kulhalli and S. C Shinde, "DNA Cryptography using Index-based Symmetric DNA Encryption Algorithm", International Journal Of Engineering Research and Technology, ISSN 0974-3154 Vol 10, No1,2017

8. A.K. Kaundal, A.K. Verma, DNA based cryptography: a review, *International Journal of Information & Computation Technology* **4**(7), 693–698 (2014). ISSN 0974-2239
9. G. Jacob, A. Murugan, DNA based cryptography: an overview and analysis. ResearchGate (2013)
10. S. Karthiga, E. Murugavalli, DNA Cryptography, *International Research Journal of Engineering and Technology*, **5**, (2018). p-ISSN 2395-0072
11. A. Das, S.K. Sarma, S. Deka, Data security with DNA cryptography, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering* 2019, 3–5 July, 2019, London, U.K., pp 246–251 (2019)

Extended Performance Research on IEEE 802.11 a WPA Multi-node Laboratory Links

J. A. R. Pacheco de Carvalho, H. Veiga, C. F. Ribeiro Pacheco, and A. D. Reis

Abstract Wireless communications, involving electronic devices, are increasingly important. Performance is a fundamental issue, leading to more reliable and efficient communications. Security is also, no doubt, most important. Laboratory measurements were achieved about several performance aspects of Wi-Fi IEEE 802.11a 54 Mbps WPA links. Our study enriches performance evaluation of this technology, using accessible equipment (HP V-M200 access points and Linksys WPC600N adapters). New accurate results are given, namely at OSI level 4, from TCP and UDP experiments. TCP throughput is measured against TCP packet length. Jitter and percentage datagram loss are evaluated versus UDP datagram size. Results are examined for point-to-point, point-to-multipoint and four-node point-to-multipoint links. Comparisons are also made mainly to related data obtained for Open links. Conclusions are extracted about performance of the links.

Keywords IEEE 802.11a · Multi-Node links · TCP packet size · UDP datagram size · Wi-Fi · Wireless network laboratory performance · WLAN · WPA2

J. A. R. P. de Carvalho (✉) · C. F. R. Pacheco · A. D. Reis
Departamento de Física, Grupo de Investigação APTEL, Universidade da Beira Interior, Covilhã
6201-001, Portugal
e-mail: pacheco@ubi.pt

C. F. R. Pacheco
e-mail: a17597@ubi.pt

A. D. Reis
e-mail: adreis@ubi.pt

H. Veiga
Centro de Informática, Grupo de Investigação APTEL, Universidade da Beira Interior, Covilhã
6201-001, Portugal
e-mail: hveiga@ubi.pt

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_14

1 Introduction

Electromagnetic waves in several frequency ranges, propagating in the air, have definitely contributed to the development of contactless communication technologies. Typical examples of wireless communications technologies are wireless fidelity (Wi-Fi) and free space optics (FSO), using microwaves and laser light, respectively. Their importance and utilization have been spreading worldwide.

Wi-Fi adopted microwave technology. Versatility, mobility and favourable prices are provided. Wi-Fi has seen its importance and utilization enlarging. It enhances traditional wired networks. Both ad hoc and infrastructure modes are used. In this second case, a wireless access point, AP, provides communications of Wi-Fi electronic devices with a wired based local area network (LAN) through a switch/router. Thus, a wireless local area network (WLAN), based on the AP, is set. At the home level, personal devices are permitted to communicate through a wireless personal area network (WPAN). Point-to-point (PTP) and point-to-multipoint (PTMP) microwave links are used in the 2.4 and 5 GHz bands, with IEEE 802.11a, 802.11b, 802.11 g, 802.11n and 802.11ac standards [1]. The increasing use of the 2.4 GHz band has led to strong electromagnetic interference. Therefore, the use of the 5 GHz band is very convenient, although absorption is larger and ranges shorter. Wi-Fi communications are not very influenced by rain or fog, as wavelengths are in the range 5.6–12.5 cm. On the contrary, rain or fog indubitably degrade FSO communications, as the typical wavelength range for the laser beam is 785–1550 nm.

Wi-Fi has nominal transfer rates up to 11 (802.11b), 54 Mbps (802.11 a, g), 600 Mbps (802.11n) and 6.9 Gbps (802.11ac). The medium access control of Wi-Fi is carrier sense multiple access with collision avoidance (CSMA/CA). 802.11a, g provide a multi-carrier modulation scheme called orthogonal frequency division multiplexing (OFDM) that allows for binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK) and quadrature amplitude modulation (QAM) of the 16-QAM and 64-QAM density types. One spatial stream (one antenna) and coding rates up to 3/4 are possible and a 20 MHz channel. 802.11a and 802.11 g work in the 5 and 2.4 GHz bands, respectively.

Studies have been published on wireless communications, wave propagation [2, 3], practical implementations of WLANs [4], performance analysis of the effective transfer rate for 802.11b PTP links [5], and 802.11b performance in crowded indoor ambiances [6].

Performance increase has been a central issue, giving more reliable and efficient communications. Requisites have been published both for traditional and new telematic applications [7].

Wi-Fi security is critically important for secretiveness reasons. Microwave radio signals travel through the air and can be quickly captured. Security methods have been developed to provide certification such as, by increasing order of safeness, wired equivalent privacy (WEP), Wi-Fi protected access (WPA) and Wi-Fi protected access II (WPA2).

Several performance measurements have been published for 2.4 and 5 GHz Wi-Fi Open [8, 9], WEP [10], WPA [11, 12] and WPA2 [13] links, as well as very high speed FSO [14]. Performance evaluation of IEEE 802.11-based Wireless Mesh Networks has been given [15]. Studies are published on modelling TCP throughput [16]. A formula that bounds average TCP throughput is available [17].

It is worthwhile investigating the effects of TCP packet size, UDP datagram size, network topology, increasing levels of security encryption, on link performance and compare equipment performance for several standards. Studies have been published for 5 GHz 802.11n WPA2 links [18]. In the present work new Wi-Fi results arise from measurements on 802.11a WPA multi-node links at 54 Mbps, namely through OSI level 4 from TCP and UDP experiments. Performance is evaluated and compared in laboratory measurements of WPA PTP, three-node point-to-multipoint (PTMP) and four-node point-to-multipoint (4 N-PTMP) links using available equipment. TCP throughput is measured against TCP packet length. Jitter and percentage datagram loss are evaluated versus UDP datagram size. In comparison to previous work [12] extended investigations on performance are realized.

In prior and actual state of the art, several Wi-Fi links and technologies have been examined. Performance evaluation has been identified as a centrally important criterion to determine communications quality. The incentive to this work is to evaluate and compare performance in laboratory measurements of WPA multi-node 802.11a links at 54 Mbps using available equipment. Thus permitting to increase the expertise about Wi-Fi (IEEE 802.11 a) link performance. The problem statement is that performance needs to be evaluated under several TCP and UDP parameterizations and link topologies under security encryption. The proposed solution uses an experimental setup and method, permitting to check signal to noise ratios (SNR) and noise levels (N), measure TCP throughput (from TCP connections) versus TCP packet size, and UDP jitter and percentage datagram loss (from UDP communications) against UDP datagram size.

The structure for the rest of the paper is as follows: Sect. 2 is about the experimental conditions i.e. the measurement setup and procedure. Results and discussion are given in Sect. 3. Section 4 presents Conclusions.

2 Experimental Details

The experiments were made during the third quarter 2019. Here a HP V-M200 access point [19] was used, with three external dual-band 3 × 3 MIMO antennas, IEEE 802.11 a/b/g/n, software version 5.4.1.0-01-16481, a 1000-Base-T/100-Base-TX/10-Base-T layer 2 3Com Gigabit switch 16 and a 100-Base-TX/10-Base-T layer 2 Allied Telesis AT-8000S/16 switch [20]. Three PCs were used having a PCMCIA IEEE.802.11 a/b/g/n Linksys WPC600N wireless adapter with three internal antennas [21], to enable 4 N-PTMP links to the access point. In every type of experiment, an interference free communication channel was used (ch. 36). This was mainly found out through a portable computer, equipped with a Wi-Fi 802.11

a/b/g/n adapter, running Acrylic WiFi software [22]. WPA encryption with TKIP was activated in the AP and the wireless adapters of the PCs, with a pre-shared key composed of twenty six hexadecimal characters. The experiments were conducted under far-field conditions. No power levels above 30 mW (15 dBm) were used, as the wireless equipments were neighbouring. The distances concerned were much larger than the wavelength used (5.8 cm).

A practical laboratory arrangement has been planned and set up for the measurements, as shown in Fig. 1. Up to three wireless links to the AP are possible. At OSI level 4, measurements were made for TCP connections and UDP communications using Iperf software [23]. For a TCP client/server connection (TCP New Reno, RFC 6582, was used), TCP throughput was collected for a given TCP packet size, varying from 0.25 to 64 k bytes. For a UDP client/server communication with a given bandwidth parameter, UDP jitter and percentage loss of datagrams were obtained for a given UDP datagram size, varying from 0.25 to 64 k bytes.

The Wi-Fi network was the following. One PC, with IP 192.168.0.2 was the Iperf server and the other PCs, with IPs 192.168.0.6 and 192.168.0.50, were the Iperf clients (client1 and client2, respectively). Jitter, which is the root mean square of differences between consecutive transit times, was constantly computed by the server, obeying to the real time protocol RTP, in RFC 1889 [24]. A control PC, with IP 192.168.0.20, was mainly used to set the configuration of the AP. The net mask was 255.255.255.0. Three types of experiments were possible: PTP (two nodes), using the client1 and the control PC as server; PTMP (three nodes), using the client1 and the 192.168.0.2 server PC; 4 N-PTMP (four nodes), using simultaneous

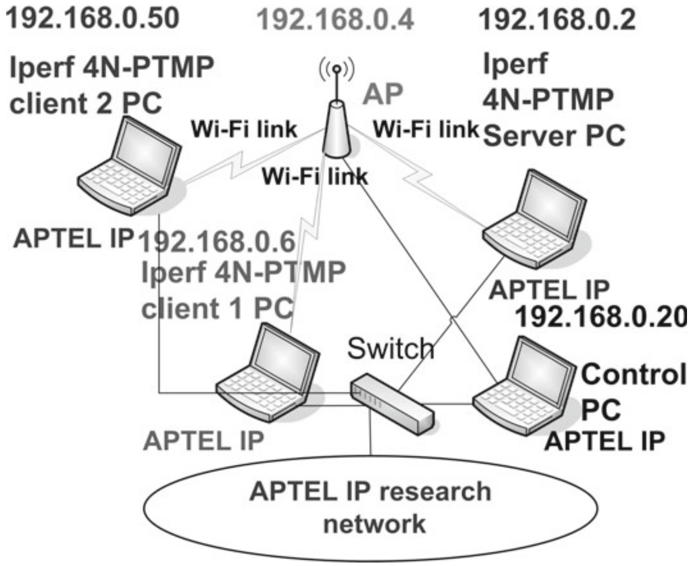


Fig. 1 Experimental laboratory setup scheme

connections/communications between the two client PCs and the 192.168.0.2 server PC.

The server and client PCs were HP nx9030 and nx9010 portable computers, respectively. The control PC was an HP nx6110 portable computer. Windows XP Professional SP3 was the operating system. The PCs were arranged to enable maximum resources to the present work. Batch command files have been re-written for the new TCP and UDP research.

The results were collected in batch mode and recorded as data files to the client PCs disks. Every PC had a second (Ethernet) network adapter, to permit remote access from the IP APTEL (Applied Physics and Telecommunications) Research Group network, via switch.

3 Results and Discussion

WPA encryption and a nominal rate of 54 Mbps were manually configured in every wireless network adapter of the PCs. Nominal transmit and receive rates were monitored in the AP along the experiments. They were regularly 54 Mbps. For every TCP packet size in the range 0.25–64 k bytes, and for every corresponding UDP datagram size in the same range, data were collected for the WPA multi-node links at OSI levels 1 (physical layer) and 4 (transport layer) using the arrangement of Fig. 1. For every link type and TCP packet size an average TCP throughput was calculated from a set of experiments. This value was taken as the bandwidth parameter for every related UDP test, giving average jitter and average percentage datagram loss.

At OSI level 1, signal to noise ratios (SNR, in dB) and noise levels (N, in dBm) were collected in the AP. Signal gives the strength of the radio signal the AP receives from a client PC, in dBm. Noise means how much background noise, due to radio interference, exists in the signal path between the client PC and the AP, in dBm. The lower the value is, the weaker the noise. SNR indicates the relative strength of client PC radio signals versus noise in the radio signal path, in dB. SNR is a good indicator for the quality of the radio link between the client PC and the AP. The collected data were similar for all types of experiments. Typical values are given in Fig. 2. The links had good, high, SNR values.

The main average TCP and UDP results are compiled in Table 1, for WPA and Open links, and every link topology (PTP, PTMP and 4 N-PTMP). The statistical analysis, including computations of confidence intervals, was done as in [25].

In Fig. 3, polynomial fits were made (shown as y versus x), using the Excel worksheet, to the TCP throughput data for WPA multi-node links, where R^2 is the coefficient of determination. It provides the goodness of fit. A value of 1.0 implies a perfect fit to data. It was found that, on average, the best TCP throughputs are for PTP, both for WPA and Open links (Table 1). In passing from PTP to PTMP, throughput reduces to 45%. In comparison to PTMP, 4 N-PTMP throughput falls to 52%. Similar trends are visible for Open links. This is due to increase of processing requirements for the AP to maintain links between PCs. No significant sensitivity

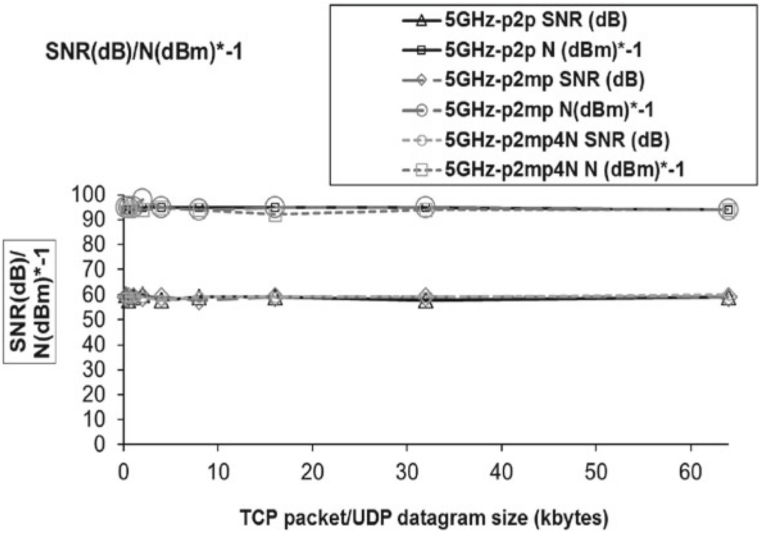


Fig. 2 Typical SNR (dB) and N (dBm). WPA links

Table 1 Average Wi-Fi IEEE 802.11a 54 Mbps WPA and open results: PTP; PTMP; 4 N-PTMP

Link type	WPA			Open		
Parameter/link type	PTP	PTMP	4 N-PTMP	PTP	PTMP	4 N-PTMP
TCP throughput (Mbps)	23.0 + −0.7	10.3 + − 0.3	5.4 + −0.2	23.1 + −0.7	10.5 + 0.3	5.4 + −0.2
UDP-jitter (ms)	4.0 + − 0.5	3.1 + −0.8	3.7 + -0.8	3.5 + − 0.2	2.6 + − 0.4	2.6 + −0.6
UDP-% datagram loss	2.5 + − 0.3	7.5 + −0.3	6.0 + -0.5	1.9 + − 0.2	7.5 + − 0.4	4.7 + −0.2

to WPA was found within the experimental error. Figure 3 evidences a fair increase in TCP throughput with packet size. For small packets, there is a large overhead, as there are small amounts of data that are sent in contrast to the protocol components. The role of the frame is very heavy in Wi-Fi. For larger packets, overhead reduces; the amount of sent data overcomes the protocol components. WPA2 802.11n results [18] show better TCP throughput performances and similar trends than WPA 802.11a data. WPA2 802.11ac TCP throughput results [13] show the best performance.

In Figs. 4 and 5, the data points representing jitter and percentage datagram loss for WPA links were joined by smoothed lines. Log 10 based scales were applied to the horizontal axes, for providing data details. Similar data are given in Figs. 6 and 7 for Open links. It was found that, on average, the best jitter performances are for PTMP, for both WPA and Open links (Table 1). This is surprising, and so far unexplained, since we would expect a degradation of jitter performance due to link

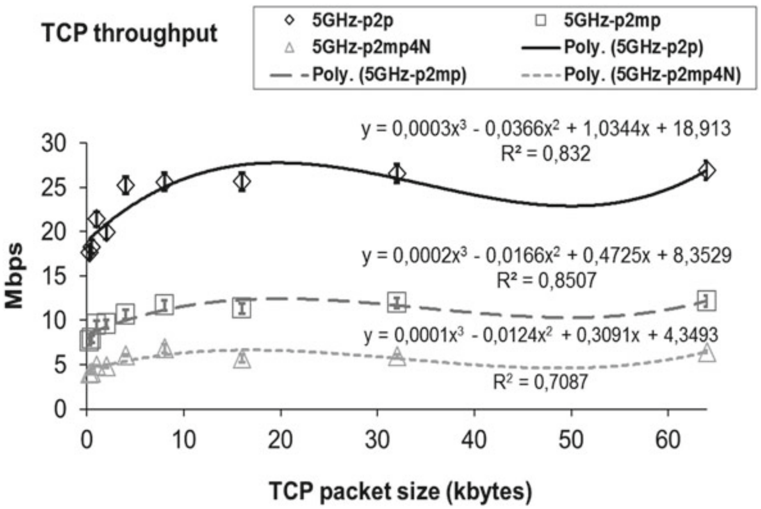


Fig. 3 TCP throughput (y) versus TCP packet size (x). WPA links

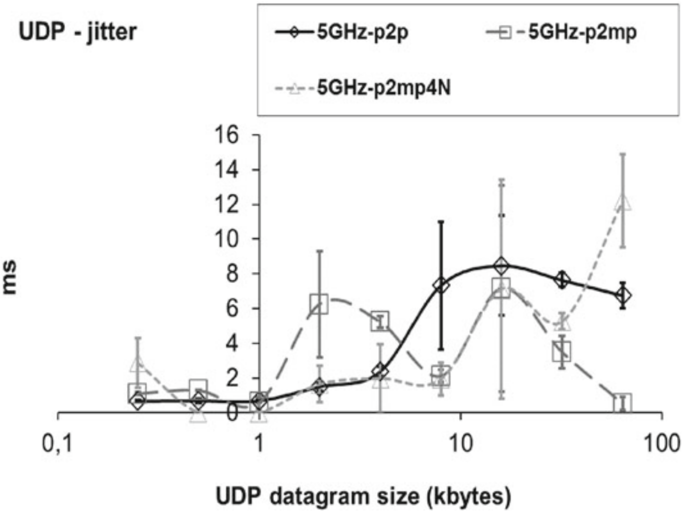


Fig. 4 UDP—jitter versus UDP datagram size. WPA links. Horizontal axis log scale

topology with increased number of nodes, where processing requirements of the AP increase for providing links between PCs. There are oscillations in the PTMP and 4 N-PTMP jitter curves (Figs. 4 and 6). For PTP and small sized datagrams, jitter is small. There are small delays in sending datagrams. Latency is also small. Jitter increases for larger datagram sizes. WPA did not show a significantly visible effect

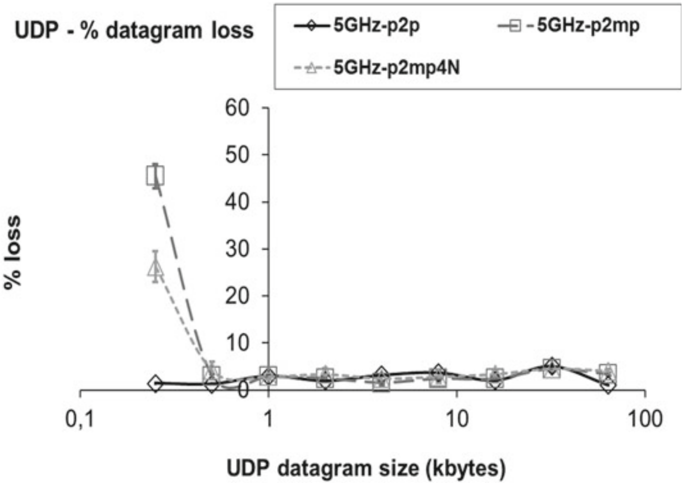


Fig. 5 UDP—percentage datagram loss versus UDP datagram size. WPA links. Horizontal axis log scale

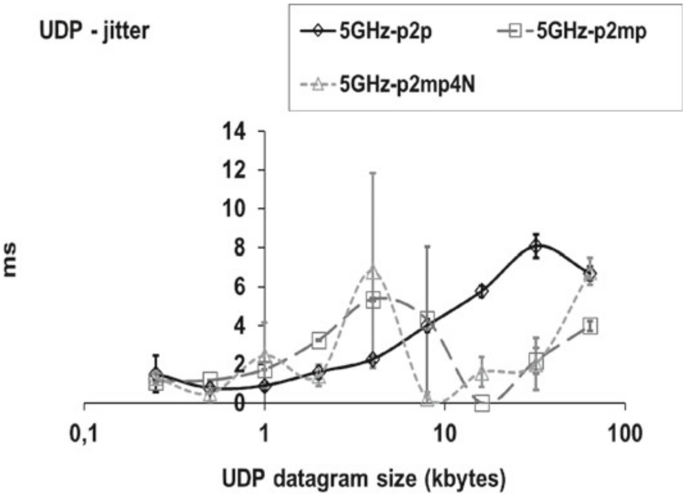


Fig. 6 UDP—jitter versus UDP datagram size. Open links. Horizontal axis log scale

on jitter performance. However, average jitter performances degrade (Table 1) in passing from Open to WPA links, where data length grows due to encryption.

Concerning average percentage datagram loss, performances were generally found, on average, to degrade due to link topology, by increasing the number of nodes (Table 1). There is an increase of processing requirements for the AP to keep links between PCs. Generally, Figs. 5 and 7 show larger percentage datagram losses

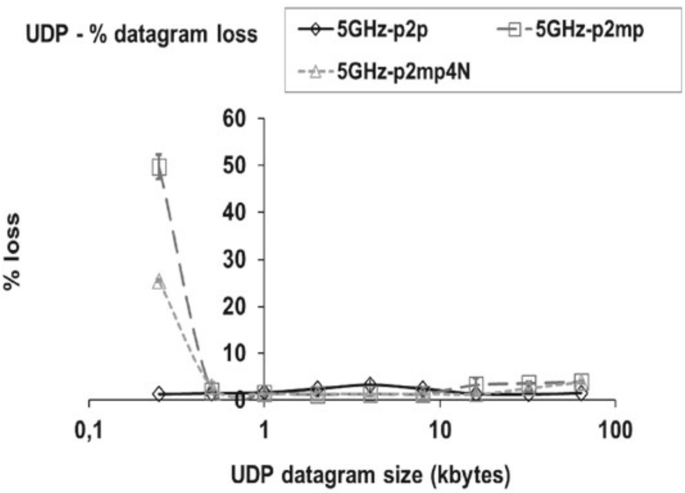


Fig. 7 UDP—percentage datagram loss versus UDP datagram size. Open links. Horizontal axis log scale

for small sized datagrams, when the amounts of data to send are small in comparison to the protocol components. There is considerable processing of frame headers and buffer management. For larger datagrams, percentage datagram loss is lower. However, large UDP segments originate fragmentation at the IP datagram level, resulting in higher losses. WPA, where data length increases due to encryption, generally shows the effect of degrading average percentage datagram loss performance for all link topologies.

TCP throughput and percentage datagram loss have generally shown performance degradations due to link topology. As CSMA/CA is the medium access control, the available bandwidth and the airtime are divided by the nodes accessing the medium. WPA has shown to degrade datagram loss performances.

Further experiments were made under similar conditions with 128-bit WEP multi-node links. It was found that TCP throughput and percentage datagram loss exhibit performance degradations due to link topology, by increasing the number of nodes. WEP, where data length increases due to encryption, was found not to diminish TCP throughput within the experimental error. However to degrade, mainly, datagram loss performance.

Present results show that 5 GHz 802.11n WPA2 [18] gives better performances than 802.11a WPA for both TCP, and (PTP and PTMP) jitter and datagram loss.

4 Conclusions

In the present work a functional laboratory setup arrangement was planned and realized, that permitted systematic performance measurements using available wireless equipment (V-M200 access points from HP and WPC600N adapters from Linksys) for Wi-Fi (IEEE 802.11 a) in WPA multi-node (PTP, PTMP and 4 N-PTMP) links.

Through OSI layer 4, TCP and UDP performances were measured versus TCP packet size and UDP datagram size, respectively. TCP throughput, jitter and percentage datagram loss were measured and compared for WPA and Open links, for all link topologies. TCP throughput was found to increase with packet size. No significant sensitivity to WPA was found within the experimental error. As for jitter, for PTP and small sized datagrams, it is found small. It increases for larger datagrams. WPA was found to degrade average jitter performance for all link topologies. Concerning percentage datagram loss, it was found high for small sized datagrams, chiefly for PTMP. For larger datagrams, it diminishes. However, large UDP segments originate fragmentation at the IP datagram level, leading to higher losses. TCP throughput and percentage datagram loss were found, generally, to show performance degradations due to link topology, in increasing the number of nodes. Processing requirements for the AP are higher, to ensure links between PCs. As CSMA/CA is the medium access control, the available bandwidth and the airtime are divided by the nodes using the medium. WPA has shown to degrade average datagram loss performances. WEP was found not to decrease TCP throughput within the experimental error. However to degrade, mainly, datagram loss performances. The present results show that 5 GHz 802.11n WPA2 gives better performances than 802.11a WPA for both TCP and (for PTP and PTMP) jitter and datagram loss.

Further performance studies are planned for several standards, equipment, topologies, security settings and noise conditions, not only in laboratory but also in outdoor environments involving, mainly, medium range links

Acknowledgements Supports from University of Beira Interior and FCT (Fundação para a Ciência e a Tecnologia)/PEst-OE/FIS/UI0524/2014(ProjectoEstratégico-UI524-2014) are acknowledged.

References

1. Web site. http://standards.ieee.org; IEEE 802.11a, 802.11b, 802.11 g, 802.11n, 802.11i standards
2. J.W. Mark, W. Zhuang, *Wireless Communications and Networking* (Prentice-Hall Inc, Upper Saddle River, NJ, 2003)
3. T.S. Rappaport, *Wireless Communications Principles and Practice*, 2nd edn. (Prentice-Hall Inc, Upper Saddle River, NJ, 2002)
4. W.R. Bruce III, R. Gilster, *Wireless LANs End to End* (Hungry Minds Inc, NY, 2002)
5. M. Schwartz, *Mobile Wireless Communications*, Cambridge University Press (2005)

6. N. Sarkar, K. Sowerby, High performance measurements in the crowded office environment: a case study, in *Proceeding of ICCT'06-International Conference on Communication Technology*, pp. 1–4, Guilin, China, 27–30 Nov (2006)
7. F. Boavida, E. Monteiro, *Engenharia de Redes Informáticas*, 10th edn. (FCA-Editora de Informática Lda, Lisbon, 2011)
8. J.A.R. Pacheco de Carvalho, H. Veiga, P.A.J. Gomes, C.F. Ribeiro Pacheco, N. Marques, A.D. Reis, Wi-Fi point-to-point links—performance aspects of IEEE 802.11 a, b, g laboratory links, in ed. by A. Sio-Iong, L. Gelman, *Electronic Engineering and Computing Technology, Series: Lecture Notes in Electrical Engineering*, vol. 60. (Springer, Netherlands, 2010), pp. 507–514
9. J.A.R. Pacheco de Carvalho, H. Veiga, C.F. Ribeiro Pacheco, A.D. Reis, Extended performance research on Wi-Fi IEEE 802.11 a, b, g laboratory open point-to-multipoint and point-to-point links, in ed. by A. Sio-Iong, Y. Gi-Chul, L. Gelman, *Transactions on Engineering Technologies*, (Springer, Singapore, 2016), pp. 475–484
10. J.A.R. Pacheco de Carvalho, C.F. Ribeiro Pacheco, A.D. Reis, H. Veiga, Extended performance studies of Wi-Fi IEEE 802.11a, b, g laboratory WEP point-to-multipoint and point-to-point links, in ed. by Y Gi-Chul, A. Sio-Iong, L. Gelman, *Transactions on Engineering Technologies: World Congress on Engineering* (Springer, Gordrecht 2015), pp. 563–572
11. J.A.R. Pacheco de Carvalho, H. Veiga, C.F. Ribeiro Pacheco, A.D. Reis, Extended performance studies of Wi-Fi IEEE 802.11a, b, g laboratory WPA point-to-multipoint and point-to-point links, in ed. by Gi-Chul Y., Sio-Iong A., L. Gelman, *Transactions on Engineering Technologies: Special Volume of the World Congress on Engineering 2013*, (Gordrecht: Springer, 2013), pp. 455–465
12. J.A.R. Pacheco de Carvalho, H. Veiga, C.F. Ribeiro Pacheco, A.D. Reis, Performance evaluation of IEEE 802.11a 54 Mbps WPA laboratory links, Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress of Engineering 2019, 3–5 July 2019, London, U.K., pp. 565–570
13. J.P. de Carvalho, H. Veiga, C.R. Pacheco, A. Reis, Performance research on IEEE 802.11 ac laboratory links, WSEAS Trans. Commun. **18**(#25), 185–190 (2019). (ISSN 1109-2742, E-ISSN 2224-2864)
14. J.A.R. Pacheco de Carvalho, N. Marques, H. Veiga, C.F.F. Ribeiro Pacheco, A.D. Reis, Performance measurements of a 1550 nm Gbps FSO link at Covilhã City, Portugal, in *Proceeding Applied Electronics 2010–15th International Conference 8–9 Sept 2010*, University of West Bohemia, Pilsen, Czech Republic, pp. 235–239
15. D. Bansal, S. Sofat, P. Chawla, P. Kumar, Deployment and evaluation of IEEE 802.11 based wireless mesh networks in campus environments, Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2011, WCE 2011, 6–8 July, 2011, London, U.K., pp. 1722–1727
16. J. Padhye, V. Firoiu, D. Towsley, J. Kurose, Modeling TCP throughput: a simple model and its empirical validation, in *Proceeding of SIGCOMM Symposium Communications, Architecture and Protocols*, Aug 1998, pp. 304–314
17. M. Mathis, J. Semke, J. Mahdavi, The macroscopic behavior of the TCP congestion avoidance algorithm. ACM SIGCOMM Comput. Commun. Rev. **27**(3), 67–82 (1997)
18. J.A.R. Pacheco de Carvalho, Hugo Veiga, Claudia F. Ribeiro Pacheco, Antonio D. Reis, Extended performance research on 5 GHz IEEE 802.11n WPA2 laboratory links, in ed. by A. Sio-Iong, L. Gelman, H. Kon Kim, *Transactions on Engineering Technologies* (Springer, Singapore, 2019), pp. 313–323. https://doi.org/10.1007/978-981-13-0746-1_24. (ISBN 978-981-13-0745-4)
19. HP V-M200 802.11n access point management and configuration guide 2010. http://www.hp.com. Accessed 3 Jan 2019
20. AT-8000S/16 level 2 switch technical data 2009. http://www.alliedtelesis.com. Accessed 10 Dec 2015
21. WPC600N notebook adapter user guide 2008. http://www.linksys.com. Accessed 10 Jan 2012
22. Acrylic WiFi software 2019. http://www.acrylicwifi.com. Accessed 8 Jan 2019
23. Iperf software (2019). http://iperf.fr. Accessed 16 Feb 2019

24. Network Working Group. RFC 1889-RTP: a transport protocol for real time applications. http://www.rfc-archive.org. Accessed 3 Jan 2019
25. P.R. Bevington, Data reduction and error analysis for the physical sciences, Mc Graw-Hill Book Company, 1969

A Low-Voltage Fourth-Order Switched-Capacitor Filter with 2.5 μm-Channel and Multi-Channel Dynamic Switching Bias OP Amplifiers

Hiroo Wakaumi

Abstract Wideband filters employing OP Amplifiers (OP Amps) are required in sensing devices. A low-voltage 4th-order Switched-Capacitor Low-Pass Filter (SC LPF) employing 2.5 μm channel-length based 3-V power Low-Voltage Folded-Cascode CMOS OP Amps with a Dynamic Switching Bias circuit (LV DSBFC OP Amps) capable of processing video signals, which enables low power consumption and operation in wide bandwidths, is proposed. Through SPICE simulations, this 4th-order SC LPF showed a suitable gain below −31.5 dB at over 5 MHz within the stop-band. Its power consumption reduced to 67% of that observed in the static operation of the OP Amps, and decreased to 56.9% of that in the 4th-order SC LPF with conventional 5-V power DSBFC OP Amps. To improve the stop-band attenuation, a Multi-Channel 4th-order SC LPF (MC SC LPF) with MC DSBFC OP Amps (MC OP Amps) consisting of 1–2 μm multi channel-length MOSFETs and short-channel sampling and CMOS switches was also presented. The MC OP Amp enables a high speed filtering operation while keeping a low distortion. Through simulations, the attenuation of the MC SC LPF decreased to below −41 dB, which is practically usable, nearly 10 dB lower than that in the 2.5 μm-channel SC LPF due to a high speed settling operation of the MC OP Amp and the great decrease of feed-through in the switches.

Keywords Switched capacitor circuit · Filter · CMOS · Operational amplifier · Dynamic switching · Video signal processing

1 Introduction

Wideband filters are essential for signal processing in video electronic appliances. Specifically, a wideband Low-Pass Filter (LPF) is needed in sensing devices such as a CCD camera handling a wide bandwidth video signal of over 2 MHz. The CMOS Switched Capacitor (SC) techniques suitable for realizing analog signal processing

H. Wakaumi (✉)
Tokyo Metropolitan College of Industrial Technology, 1-10-40 Higashi-Ooi, Shinagawa, Tokyo 140-0011, Japan
e-mail: waka781420j@ab.auone-net.jp

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_15

ICs, have promising use in video signal bandwidth circuits. It has been demonstrated that SC techniques using CMOS Operational Amplifiers (OP Amps) are useful for implementing analog functions such as filters [1–4]. However, the use of several OP Amps results in large power consumption and may cause unstable operation. Until now, several approaches to decrease the power consumption of OP Amps have been considered, including the development of ICs that work at low power supply voltages [5]. A clocked current bias scheme for folded-cascode OP Amps suitable for achieving a wide dynamic range has been typically proposed to decrease the power consumption of the OP Amp itself [6]. Because the circuit requires complicated four-phase bias-current control pulses and bias circuits, it is not suitable for high speed operation. A high-gain Folded-Cascode (FC) OP Amp using identical 0.8 μm channel-length MOSFETs and a 5-V power supply, has been designed [7]. Although its power dissipation is lower than 5 mW, this circuit using resistances for current mirrors is not suitable for realizing as an IC. An FC OP Amp using quasi-floating gate FETs consisting of identical channel length of 0.24 μm (except resistors) has been also proposed for high speed operation keeping low power dissipation [8]. Due to the low power supply of 1 V, its slew rate is not so high (19 V/μs) and distortion is not necessarily low enough (0.5% for a small output signal of 0.45 Vp-p).

Recently, the author proposed an FC CMOS OP Amp with a Dynamic Switching Bias circuit (DSBFC OP Amp), of simple configuration, to provide low power consumption while maintaining a high speed switching operation suitable for processing video signals [9]. The power consumption of this OP Amp with the 5-V power supply voltage was not necessarily low enough for use in low-voltage signal processing applications. Then, the author proposed a Low-Voltage DSBFC OP Amp (LV DSBFC OP Amp) consisting of 2.5 μm channel-length MOSFETs (except the DSB circuit part) with the 3-V power supply voltage to provide lower power consumption and showed its practicability on a 2nd-order SC LPF using this OP Amp [10]. However, its SC LPF has a problem of an insufficient roll-off gain characteristic. Though the availability of a 4th-order SC LPF with 5-V power DSBFC OP Amps was studied previously [11], its power dissipation was not low enough. Under the recent progress of miniaturization of equipment, the development of a practicable low-power LPF with low-voltage OP Amps is expected.

The gain attenuation in a stop-band is determined by the operation speed of the OP Amp and the holding characteristic of sampling and CMOS switches. For the high speed operation of OP Amps, the channel length of MOSFETs constituting OP Amps needs to be short as much as possible. Considering the easiness of design of these circuits, an identical channel length for OP Amps using short channel lengths close to 1 μm is used [7, 8, 12]. However, their distortion increases due to the channel modulation effect of such short channel MOSFETs. In order to minimize this distortion, the channel length optimization of MOSFETs in constituent elements becomes an important issue. Such approach for short channel FETs in OP Amps has not been reported until now.

In this paper, a low-voltage 4th-order SC LPF employing 2.5 μm-channel LV DSBFC OP Amps with the 3-V power supply voltage [13], which enables lower power consumption and is suitable for achieving wide bandwidths IC due to the sharp roll-off characteristic and low power supply voltage operation, is proposed. Its configuration under the optimization of sample-hold circuit and OP Amp load capacitance is shown in Sect. 2. The availability of this SC LPF for the performance of frequency response and power dissipation is evaluated in Sect. 3. Then, the design approach of a Multi-Channel DSBFC OP Amp (MC OP Amp) consisting of multi channel-length MOSFETs which enables a high speed settling and low distortion, is described in Sect. 4. To achieve the practicable stop-band attenuation characteristics, a Multi-Channel 4th-order SC LPF (MC SC LPF) with this MC OP Amp and short-channel sampling and CMOS switches is presented in Sect. 5. The performance of the MC SC LPF is evaluated in Sect. 6. Finally, conclusions are summarized in Sect. 7.

2 Low-Voltage 4th-Order SC LPF Design

As a low-voltage 4th-order SC LPF, an IIR (Infinite Impulse Response) SC LPF with the Butterworth frequency characteristic employing 3-V power LV DSBFC OP Amps was designed. The high filter order of the fourth was chosen because it is expected to achieve a sharp roll-off gain characteristic. This 4th-order SC LPF was designed to achieve a gain of below −30 dB at over 4 MHz within the stop-band. The other design condition was set as follows. That is, a sampling frequency fs = 14.3 MHz corresponding to four times of NTSC color subcarrier frequency 3.58 MHz and a cutoff frequency fc = 2 MHz were chosen for this SC LPF, that enable it to process video signals. Under such condition, the theoretical Butterworth discrete-time transfer function obtained by bilinear transformation from s-region to z-region is given using the z-transform by (1).

$$\mathrm{H(z)} = \frac{-0.10573\left(1+\mathrm{z}^{-1}\right)^2}{1-0.74578\mathrm{z}^{-1}+0.16869\mathrm{z}^{-2}} \cdot \frac{-0.13976\left(1+\mathrm{z}^{-1}\right)^2}{1-0.98582\mathrm{z}^{-1}+0.54484\mathrm{z}^{-2}} \qquad (1)$$

The circuit configuration realizing this transfer function is shown in Fig. 1 [11]. Its operation waveforms are shown in Fig. 2. The 4th-order SC LPF was designed referencing a SC biquadratic circuit with integrators [14]. This SC LPF consists of two-stage biquadratic circuits cascading two 2nd-order SC LPFs of LPF1 and LPF2, provided with a sample-hold circuit with a holding capacitor Cs_1 and a sampling switch controlled by φ_{SH}, CMOS switches φ_1 andφ_2, capacitors A_1-E_1, G_1, I_1, A_2-E_2, G_2, and I_2, and four LV DSBFC OP Amps (OP Amp 11, OP Amp 12, OP Amp 21, and OP Amp 22). Based on the smallest coefficients of this function corresponding to a reference capacitance of 0.5 pF, each capacitance in the SC LPF was set in proportion to the reference capacitance as shown in Fig. 1 [11]. The transfer function of this SC LPF circuit except for the sample-hold circuit is identical to (1). Regarding the sample-hold circuit, the zero-order hold function due to a sample-hold effect, is multiplied to (1). The power supply voltage of V_{DD} and V_{SS} is equal to 1.5 V.

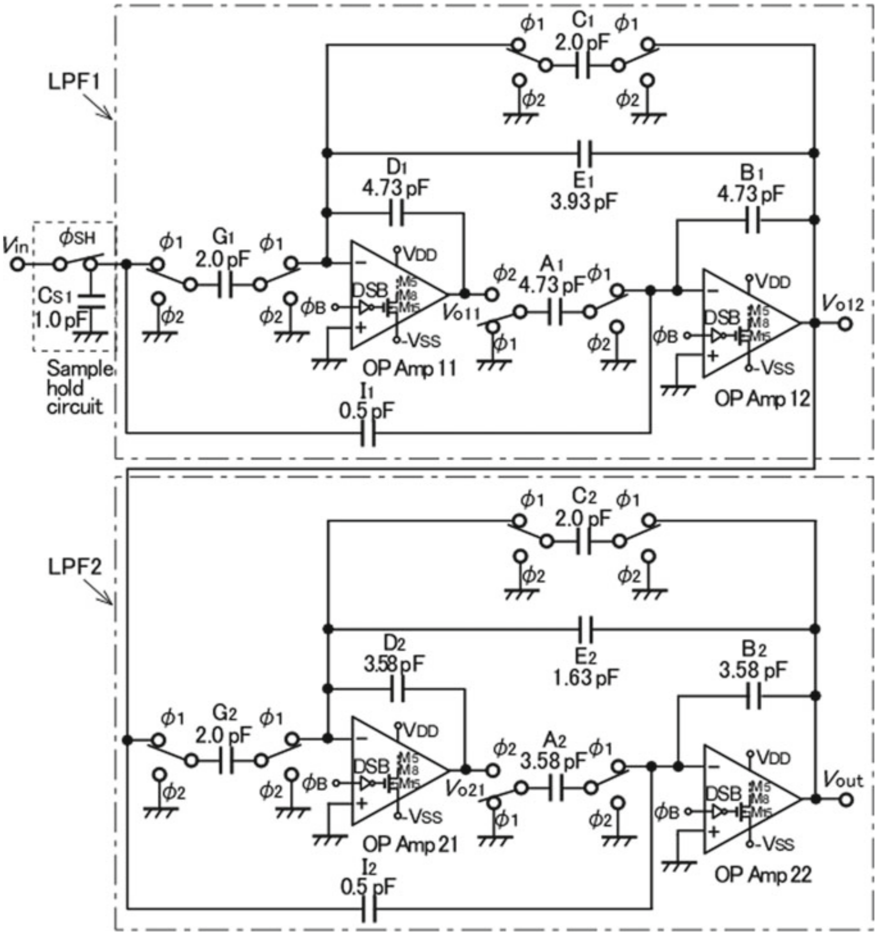
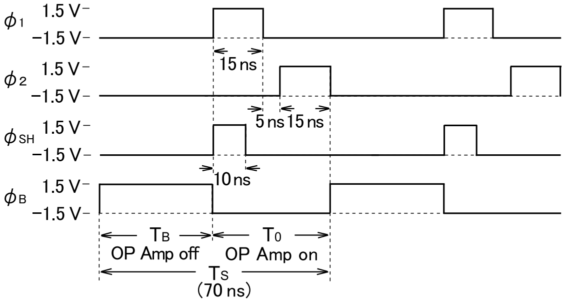


Fig. 1 Configuration of the 4th-order SC LPF

Fig. 2 Operation waveforms of the 4th-order SC LPF



In the 4th-order SC LPF, a configuration of the LV DSBFC OP Amp enabling low power consumption is shown in Fig. 3. With respect to the LV DSBFC OP Amp, the same CMOS FETs channel width/channel length W/L as that in the LV DSBFC OP Amp (2.5 μm-channel OP Amp) shown in the reference [10] was employed. Typical performances of this OP Amp are shown in Table 1. The load capacitance of this OP Amp was designed considering its switching speed and dynamic off swing as follows. Figure 4 shows gain versus load capacitance of the LV DSBFC OP Amp. The gain of the SC LPF became minimum at a load capacitance of the LV DSBFC OP Amps $C_L = 2$ pF. When C_L is smaller than 1.5 pF, the gain corresponding to the input signal frequency $f$in = 5 MHz within the stop-band deteriorates due to the mismatch of top and bottom signal levels of dynamic off swing. On the contrary, when C_L is larger than 2.5 pF, the gain deteriorates due to the slow switching speed of OP Amps. That is, C_L in the 4th-order SC LPF is optimized to 2 pF.

Fig. 3 Configuration of the low-voltage DSBFC OP Amp

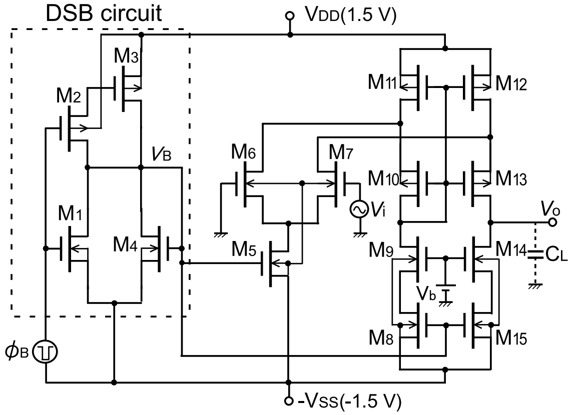


Table 1 Typical performances of 3-V power LV DSBFC OP Amp (2.5 μm-channel OP Amp). $C_L = 2$ pF

Performance parameters	3-V power LV DSBFC OP Amp
Power supply voltages	±1.5 V
Switching frequency *fs*	14.3 MHz
Open loop gain Ao	47.1 dB
Phase margin θ	39.0°
Unity gain frequency *fu*	402.1 MHz
Slew rate SR ($C_L = 10$ pF)	131 V/μs
Fall settling time t_s ($C_L = 10$ pF)	17.9 ns
Distortion THD (*fin* = 10 kHz, $V_o = 0.6\ V_{p\text{-}p}$)	0.73%
Power dissipation at 50% switching duty ratio	9.3 mW

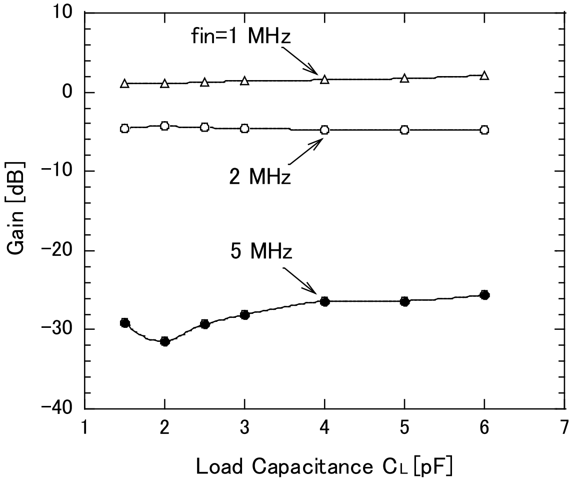


Fig. 4 Gain versus load capacitance of the LV DSBFC OP Amp

The sample-hold circuit in this SC LPF was designed as follows considering the feed-through phenomenon. Figure 5 shows gain of the 4th-order SC LPF in the DSB mode of the LV DSBFC OP Amp versus channel width Wsh of each of p-MOSFETs and n-MOSFETs in the sampling switch. The gain of the 4th-order SC LPF showed a minimum value at a Wsh of nearly 140 μm when *f*in is equal to 5 MHz within the stop-band, while its gain remains almost unchanged for *f*in of 1 and 2 MHz within the passband. When Wsh is larger than 140 μm, the feed-through caused by the difference of capacitive coupling between gate and output terminals of the above MOSFETs does not become negligible at the off-state transition of the sampling switch and so the gain corresponding to *f*in = 5 MHz increases. When Wsh is smaller than this value, a driving ability of the sampling switch becomes insufficient, which brings

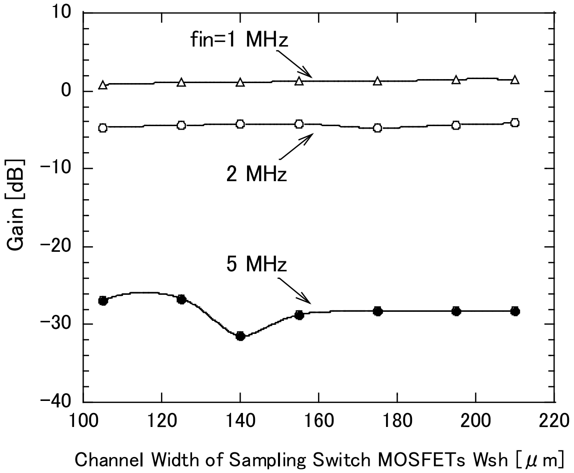
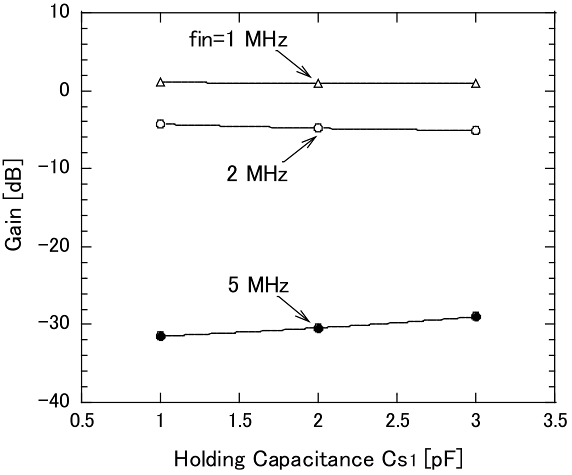


Fig. 5 Gain versus channel width of sampling switch MOSFETs. $Cs_I = 1$ pF

Fig. 6 Gain versus holding capacitance. W/L of sampling switch MOSFETs = 140/2.5 (μm/μm)



about an increase of the gain. Like this, Wsh of the sampling switch is optimized to 140 μm. That is, the sampling switch was designed to W/L = 140/2.5 (μm/μm) for each of p-MOSFETs and n-MOSFETs. The feed-through in the sample-hold circuit is also dependent on a holding capacitance. Figure 6 shows gain of the 4th-order SC LPF in the DSB mode of the LV DSBFC OP Amp versus holding capacitance. As the holding capacitance Cs_1 increases, the gain corresponding to $f$in = 5 MHz in the stop-band deteriorates little by little due to the need of a long transition time for sampling. That is, we can see that a smaller capacitance is desirable as Cs_1. So, the Cs_1 of 1 pF in this 4th-order SC LPF was also chosen.

Other CMOS switches were designed to W/L = 75/2.5 (μm/μm), which is the same one as that in the 2nd-order SC LPF [10]. CMOS switches are turned on and off by non-overlapping two-phase clock pulses φ_1 and φ_2, swinging from −1.5 V to 1.5 V. These sampling and CMOS switches were designed to have a balanced structure with each equal L and W of p-MOS and n-MOS FETs to suppress a feed-through phenomenon as much as possible. This phenomenon is easy to be caused by a capacitive coupling between gate and output terminals.

Major CMOS process parameters are given as a gate insulating film thickness $\mathrm{t_{ox}} = 50$ nm, a p-MOSFET threshold voltage $\mathrm{V_{TP}} = -0.6$ V, and an n-MOSFET threshold voltage $\mathrm{V_{Tn}} = 0.6$ V.

In this SC LPF, charge transfer operations through clock pulses φ_1, φ_2, are performed during the on-state period To of the LV DSBFC OP Amps (when the control pulse φ_B is set at −1.5 V). The off-state period $\mathrm{T_B}$ (the remaining period of the one cycle period Ts) is separately provided to realize low power consumption for this SC LPF. An input signal Vin is sampled during the sampling phase of φ_{SH} (10 ns) in the on-state period To. After sampling operation, its corresponding charge is stored on the holding capacitor Cs_1. The voltage at the off-state transition of φ_{SH} is transferred to an output terminal Vout, charging all capacitors, during the clock phase of φ_1. During the subsequent clock phase of φ_2, each charge of capacitors C_1,

C_2, G_1 and G_2 is discharged and each charge of remaining capacitors is redistributed. During such on-state period of To, the LV DSBFC OP Amps turn on by setting a bias voltage of V_B at an appropriate level of nearly 0 V and operate normally as operational amplifiers.

Subsequently, φ_B becomes 1.5 V at the off-state transition of the OP Amps, while φ_2 is switched off. During this off period T_B, the OP Amps turn off and so these do not dissipate power at all. Therefore, when T_B is set relatively long as compared to the one-cycle period Ts, the power consumption of the SC LPF is expected to become lower than that observed in ordinary static operation for an SC LPF using conventional OP Amps.

3 Low-Voltage 4th-Order SC LPF Simulation Results

The performance of the SC LPF was tested by the B^2 Spice package. The model used is MOS2. Operation waveforms for an input signal of 1 MHz with an amplitude of 0.3 V and an output load of 2 pF are shown in Fig. 7. In the 4th-order SC LPF, the output signal amplitude nearly close to the input signal was also obtained for passband frequency signals. The frequency response of the 4th-order SC LPF in the dynamic switching operation of the LV DSBFC OP Amps compared to that of the 2nd-order SC LPF is shown in Fig. 8. The roll-off characteristic in near 3–4 MHz was greatly improved compared to that in the 2nd-order SC LPF. The response was near the theoretical one from 100 kHz up to near 4 MHz. At 4 MHz within the stop-band, the gain below −28 dB was obtained. In the high frequency range over 5 MHz within the stop-band, although it deteriorated due to a sampling phase effect, the gain below −31.5 dB (a suitable level) was achieved. In this way, a wide stop-band with a high attenuation (a sharp roll-off characteristic) in the high frequency response became possible due to the two-stage biquadratic SC LPF configuration with the filter order of the fourth. Thus, it is clear that the LV DSBFC OP Amp is also applicable to the high-order SC LPF.

Power dissipation versus OP-Amp switching duty ratio in the 4th-order SC LPF with 3-V power LV DSBFC OP Amps compared to that in the 4th-order one with conventional 5-V power DSBFC OP Amps is shown in Fig. 9. The power dissipation

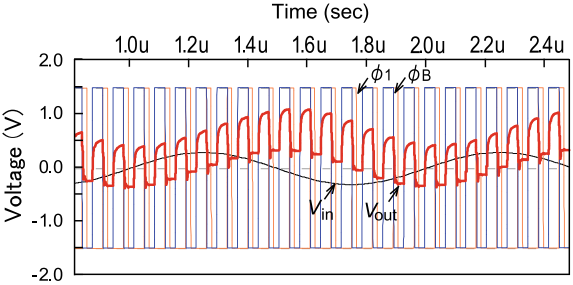


Fig. 7 Simulation waveforms for the 4th-order SC LPF. $V_{in} = 0.3\ V_{0\text{-}p}$, $f_{in} = 1$ MHz, $C_L = 2$ pF

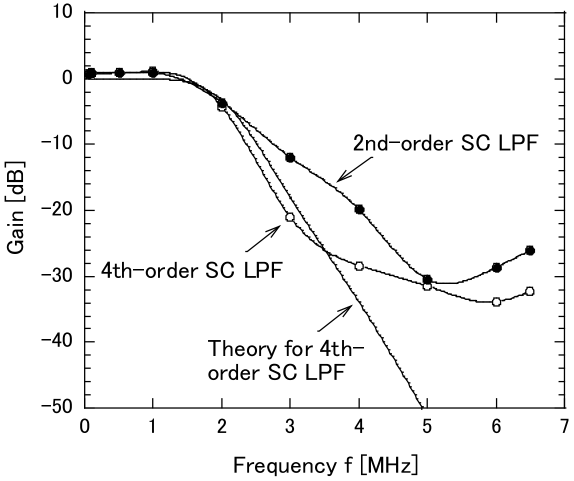


Fig. 8 Frequency response of the 4th-order and 2nd-order SC LPFs in the DSB mode of the LV DSBFC OP Amps. $T_B = 35$ ns

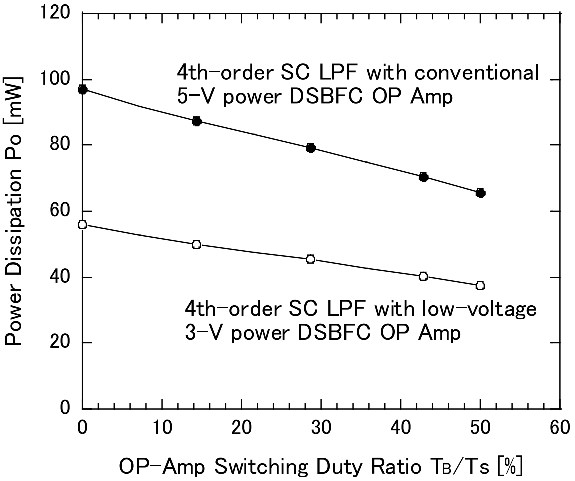


Fig. 9 Power dissipation versus OP Amp switching duty ratio in the 4th-order SC LPFs. $f_{in} = 1$ MHz

of the 4th-order SC LPF with the LV DSBFC OP Amps decreased in proportion to the off-state period T_B of the OP Amps. In the dynamic switching operation mode of $T_B = 35$ ns (= 50% switching duty ratio), the power consumption of this 4th-order SC LPF (37.4 mW) decreased to 66.8% as compared to that in the static operation of the LV DSBFC OP Amps (56 mW). This value is twice as large as that in the 2nd-order SC LPF with the LV DSBFC OP Amps (18.7 mW) because the 4th-order SC LPF consists of a cascade connection of two 2nd-order SC LPFs. However, the power consumption of this 4th-order SC LPF with the LV DSBFC OP Amps at 50% switching duty ratio was reduced to 56.9% compared to that in the 4th-order SC LPF with conventional 5-V power DSBFC OP Amps (65.7 mW). Such low power characteristic was realized by the low power supply voltages and dynamic switching

Table 2 Typical performances for the 4th-order and 2nd-order SC LPFs employing the LV DSBFC OP Amps ($V_{DD} = V_{SS} = 1.5$ V)

Performance parameters	Simulation results	
	4th-order SC LPF-this work	2nd-order SC LPF
Sampling and switching frequency *f*s	14.3 MHz	14.3 MHz
Input signal amplitude	0.3 $V_{0\text{-}P}$	0.3 $V_{0\text{-}P}$
Cutoff frequency *f*c	2 MHz	2 MHz
Gain at a stop-band over 5 MHz	≦−31.5 dB (Dynamic mode)	≦−26.0 dB (Dynamic mode)
Power consumption	56.0 mW (Static)	28.0 mW (Static)
	37.4 mW (Dynamic: $T_B/T_S = 50\%$)	18.7 mW (Dynamic: $T_B/T_S = 50\%$)

operation. Typical characteristics of the 4th-order SC LPF compared with those of the 2nd-order SC LPF employing the LV DSBFC OP Amps are listed in Table 2.

4 Design of Multi-Channel OP Amp

To improve the SC LPF performance in the stop-band region, the channel-length shortening of MOS FETs in an OP Amp is considered. A configuration of the 3-V power MC OP Amp with a DSB circuit is the same as that in Fig. 3 except the channel size and the gate oxide film thickness of MOSFETs. This OP Amp consists of a DSB circuit suitable for low power consumption and an FC OP Amp to achieve a wide dynamic range even in low-power supply voltages. As a distinctive feature, the MC OP Amp except the DSB circuit consists of multi channel-length FETs with a channel length L of 1–2 μm and a gate oxide film of 25 nm (Table 3). There is a fear that the shortening of MOSFETs channel-length might usually cause distortion. So, this time, to minimize the distortion of this OP Amp, 2 μm channel-length MOSFETs M_6, M_7, M_{10}, and M_{13} in the differential pair and current mirrors, were set. The gain Av of OP Amp is influenced directly by these MOSFETs as follows.

$$\mathrm{Av} = g_{m7} \cdot [(r_{d15} \cdot g_{m14} \cdot r_{d14})//\{(r_{d7}//r_{d12}) \cdot g_{m13} \cdot r_{d13}\}] \tag{2}$$

Here, g_{m7}, g_{m13}, and g_{m14} are a mutual conductance of MOSFETs M_7, M_{13}, and M_{14}, respectively. And r_{d12}, r_{d13}, r_{d14}, and r_{d15} are drain resistances of MOSFETs M_{12}, M_{13}, M_{14}, and M_{15}, respectively. By choosing the above long channel length for these differential pair and current mirrors, each channel modulation effect, that is, distortion caused by change of g_m, can be suppressed smaller. For other MOSFETs,

Table 3 Multi-Channel DSBFC OP Amp designed values compared with 2.5 μm-channel OP Amp. W/L = Channel width/Channel length

	Multi-Channel OP Amp	2.5 μm-Channel OP Amp
FETs	W/L [μm/μm]	W/L [μm/μm]
M_1	4.5/1.5	15/2.5
M_2	9/1.5	30/2.5
M_3	25/4	50/4
M_4	22/6	44/6
M_5	42.5/1.5	187/2.5
M_6, M_7	678/2	2000/2.5
M_8, M_{15}	32/2	92/2.5
M_9, M_{14}	69/1	1055/2.5
M_{10}, M_{13}	678/2	2000/2.5
M_{11}, M_{12}	89/1.5	390/2.5
Gate oxide film	Thickness t_{ox} [nm]	Thickness t_{ox} [nm]
	25	50

current sources of M_5, M_9 and M_{14}, and a current mirror of M_{11} and M_{12}, which do not have influence on the gain performance, have been configured using 1 or 1.5 μm channel-length MOSFETs. By constituting the OP Amp like this, a THD of 0.8% close to the conventional one (0.73%) can be obtained due to the improvement of linearity (Table 4). In the MC OP Amp, a parasitic capacitance also decreases due to the use of short-channel MOSFETs while it is pretty large due to the use of large W/L (2000 μm/2.5 μm) of M_7 and M_{13} in the 2.5 μm-channel LV DSBFC OP Amp. As a result, the slew rate SR increases and the fall settling time ts is also improved to 12 ns from the 2.5 μm-channel one of 18 ns. C_L of 10 pF for these SR and ts was chosen as an ordinary condition.

Table 4 Performances of Multi-Channel DSBFC OP Amp. $C_L = 2$ pF

Performance parameters	Multi-Channel OP Amp
Power supply voltages	±1.5 V
Switching frequency *fs*	14.3 MHz
Open loop gain Ao	47.2 dB
Phase margin θ	45.8°
Unity gain frequency *fu*	546.2 MHz
Slew rate SR ($C_L = 10$ pF)	152.8 V/μs
Fall settling time t_s ($C_L = 10$ pF)	12 ns
Distortion THD (fin = 10 kHz, Vo = 0.6 Vp-p)	0.81%
Power dissipation (50% switching duty ratio)	9.6 mW

CMOS process parameters for the MC OP Amp are identical to that in the LV DSBFC OP Amp except the gate oxide film thickness.

5 Design of Multi-Channel SC LPF

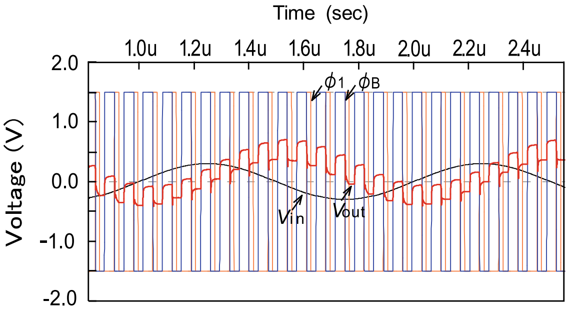
Employing the MC OP Amp, an MC 4th-order IIR SC LPF with the Butterworth frequency characteristic was designed. This SC LPF was designed to achieve a sharp roll-off characteristic with a gain of below −40 dB at over 5 MHz within the stop-band. The circuit configuration realizing this SC LPF is given by one shown in Fig. 1, replacing four OP Amps by MC OP Amps. Its operation waveforms are the same as one shown in Fig. 2. The transfer function of this SC LPF except for the sample-hold circuit is identical to (1). Regarding the sample-hold circuit, the zero-order hold function due to a sample-hold effect, is multiplied into (1). In the MC SC LPF, MOSFETs with a short L of 1.5 μm were used for the sampling switch and CMOS switches to reduce a feed-through due to a capacitive coupling between their gates and output terminals. The channel width Wsh of sampling switch MOSFETs was set at an optimum value of 78 μm in which both conditions of the small feed-through and the sufficient driving ability involved in a trade-off are satisfied. The other CMOS switches also operate with a holding function like the sample-hold circuit. The CMOS switch channel width W_B was optimized to 9 μm similarly.

The load capacitance C_L of the MC OP Amp was designed to an optimal value 2 pF considering its switching speed and dynamic off swing at the off-state transition.

6 MC SC LPF Simulation Results

Operation waveforms for a 1 MHz input signal are shown in Fig. 10. The output signal nearly close to the passband input signal was obtained. The frequency response of the MC SC LPF compared to that of the 2.5 μm-channel 4th-order SC LPF with the 2.5 μm-channel LV DSBFC OP Amp is shown in Fig. 11. The response was near

Fig. 10 Simulation waveforms for the MC SC LPF. $V_{in} = 0.3\ V_{0\text{-}p}$, $f_{in} = 1$ MHz



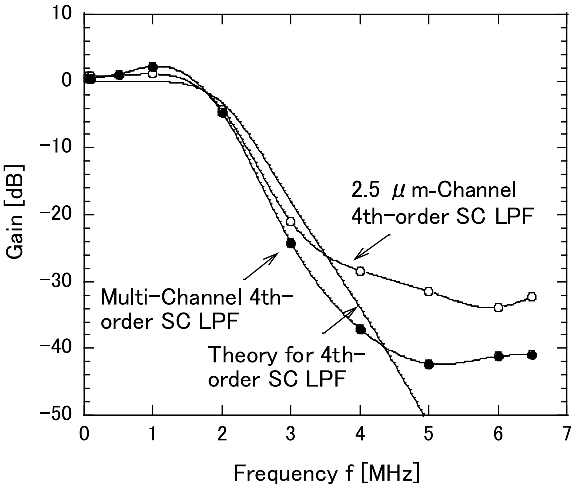


Fig. 11 Frequency response of the MC and 2.5 μm-channel 4th-order SC LPFs. $f_s = 14.3$ MHz, $T_B/T_s = 50\%$

the theoretical one from 100 kHz up to nearly 4 MHz although a gain of 2.2 dB a little larger than 0 dB for the f = 1 MHz input signal within the passband was observed. In the high frequency range over 5 MHz within the stop-band, although it deteriorated due to a sampling phase effect, a practical gain below −41 dB was achieved. This gain is considerably smaller than −31.5 dB in the SC LPF employing 2.5 μm-channel OP Amps. This is due to the marked decrease of feed-through in the sample-hold circuit (from 22 to 11 mV for input signals of −0.3 to 0.3 V) & in CMOS switches (from 17 to 3 mV) and a high speed settling operation of the MC OP Amp. Like this, a practical high attenuation in a wide stop-band became possible due to the configuration of the multi channel-length MOSFET based OP Amp and 1.5 μm channel-length sampling/CMOS switches.

The power dissipation in the MC SC LPF is nearly equal to that in the 2.5 μm-channel one as expected (Table 5).

Table 5 Typical performances of the Multi-Channel 4th-order SC LPF

Performance parameters	Simulation results (MC SC LPF)
Power supply voltages	±1.5 V
Switching frequency fs	14.3 MHz
Input signal amplitude	0.3 $V_{0\text{-}P}$
Cutoff frequency fc	2 MHz
Gain at a stop-band over 5 MHz	≦−40.8 dB (Dynamic mode)
Power consumption	56.2 mW (Static)
	38.4 mW (Dynamic: T_B/T_S = 50%)

7 Conclusions and Future Work

A 4th-order Switched-Capacitor Low-Pass Filter (SC LPF) employing 2.5 μm-channel 3-V power Low-Voltage Folded-Cascode CMOS OP Amplifiers with a Dynamic Switching Bias circuit (LV DSBFC OP Amps) capable of processing video signals was evaluated through SPICE simulations. It was confirmed that the SC LPF configuration with the increased filter order of the fourth and the 3-V power LV DSBFC OP Amp is useful for achieving a wide stop-band with a high attenuation and low power consumption. The dynamic charging operations during the on-state period of the OP Amps and non-charging operation of them during their off-state period are also useful for high speed operation, and greatly reducing the power consumption of the SC LPF. Further, it was confirmed that a Multi-Channel 4th-order Switched-Capacitor Low-Pass Filter (MC SC LPF) with MC Dynamic Switching Bias Folded-Cascode OP Amplifiers consisting of 1–2 μm multi channel-length MOSFETs is useful for achieving a wide stop-band with a practically usable attenuation maintaining low distortion and power consumption. This multi-channel configuration should be useful for the realization of not only such video signal processing filters but also low-distortion, low-power, wide-band signal processing ICs.

Furthermore, the passband response in an SC LPF is expected to be improved by the revision of a dynamic switching technique of OP Amps.

References

1. R. Gregorian, W.E. Nicholson, CMOS switched capacitor filters for a PCM voice CODEC. IEEE J. Solid-State Circ. **SC-14**(6), 970–980 (1979)
2. R. Dessoulavy, A. Knob, F. Krummenacher, E.A. Vittoz, A synchronous switched capacitor filter. IEEE J. Solid-State Circ. **SC-15**(3), 301–305 (1980)
3. A. Iwata, H. Kikuchi, K. Uchimura, A. Morino, M. Nakajima, A single-chip codec with Switched-Capacitor filters. IEEE J. Solid-State Circ. **SC-16**(4), 315–321 (1981)
4. J.-T. Wu, Y.-H. Chang, K.-L. Chang, 1.2 V CMOS Switched-Capacitor circuits, in *1996 IEEE International Solid-State Circuits Conference Digest of Technical Papers (42nd ISSCC)*, pp. 388–389, 479
5. Z. Kun, W. Di, L. Zhangfa, A high-performance folded cascode amplifier, in *2011 International Conference on Computer and Automation Engineering (ICCAE 2011)*, vol. 44 (2012), pp. 41–44
6. D.B. Kasha, W.L. Lee, A. Thomsen, A 16-mW, 120-dB linear Switched-Capacitor Delta-Sigma modulator with dynamic biasing. IEEE J. Solid-State Circ. **34**(7), 921–926 (1999)
7. Er. Rajni, Design of high gain folded-cascode operational amplifier using 1.25 um CMOS technology. Int. J. Sci. Eng. Res. **2**(11), 1–9 (2011)
8. J.M. Algueta-Miguel, A. Lopez-Martin, M.P. Garde, C.A. De la Cruz, J. Ramirez-Angulo, ±0.5 V 15 μW recycling folded cascode amplifier with 34767 MHz • pF/mA FOM. IEEE J. Solid-State Circ. Lett. **1**(7), 170–173 (2018)
9. H. Wakaumi, A folded-cascode OP amplifier with a dynamic switching bias circuit. Eng. Lett. **23**(2), 92–97 (2015)
10. H. Wakaumi, A low-voltage folded-cascode OP amplifier with a dynamic switching bias circuit, in *The Ninth International Conference on Sensor Device Technologies and Applications (SENSORDEVICES 2018)*, Sept. 2018, pp. 60–64

11. H. Wakaumi, A Fourth-Order Switched-Capacitor Low-Pass filter with dynamic switching bias OP amplifiers, in *6th International Conference on Advanced Technology & Sciences (ICAT'Riga)*, Sept. 2017, pp. 147–151
12. P.E. Allen, D.R. Holberg, *CMOS Analog Circuit Design*, 3rd edn. (Oxford, New York, NY, USA, 2012), pp. 261–342
13. H. Wakaumi, A low-voltage Fourth-Order Switched-Capacitor filter with dynamic switching bias OP Amps, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, 3-5 July, 2019, London, UK, pp. 167–171
14. T. Takebe, A. Iwata, N. Takahashi, H. Kunieda, *Switched Capacitor Circuit* (Gendai Kohgakusha, Tokyo, 2005), pp. 60–78

Reliability Centered Maintenance: Case Studies

Abdulaziz A. Bubshait and Alawi Basurrah

Abstract This paper aims to clarify the importance of applying the Reliability Centered Maintenance (RCM) methodology on critical systems. The RCM is a well-established methodology in determining and optimizing preventive maintenance strategies. The chapter shows the importance of having reliable systems in process plants through presenting case studies. The number of maintenance orders before and after the implementation of RCM has been remarkably reduced. The associated costs and equipment availability were also improved. The economic benefits, the impact of applying RCM, and cost optimization were also presented. Appling reliability tools (RCM) on plant's critical systems enable organization to ensure the reliability and availability of equipment in order to achieve the annual production target.

Keywords Equipment availability · Failure mode · Maintenance cost · Maintenance strategy · Reliability · RCM · System availability

1 Introduction

Planning and performing of preventive maintenance (PM) of technical systems is a challenging task. A tradeoff has to be made between the frequency and extension of maintenance, and costs. Preventive maintenance is used to avoid occurrence of system failures, production interruption, and reduce potential consequences of failures. Maintenance, however, could in some cases introduce failures. Both contradicting aspects are of relevance to preventive maintenance planning. Several tools and methodologies have been introduced to support planning of PM. This chapter

A. A. Bubshait (✉) · A. Basurrah
Engineering Management, King Fahd University of Petroleum & Minerals, Dhahran, Saudi Arabia
e-mail: bushait@kfupm.edu.sa

A. Basurrah
e-mail: g200429260@kfupm.edu.sa

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_16

addresses Centered Maintenance (RCM) , which is an effective and efficient methodology to reduce failures. The main objective of RCM is to reduce maintenance costs and increase reliability and safety [1, 2].

The impact of RCM implementation on critical systems of chemical plants were measured and analyzed. This chapter reports the findings of the implementation study. Four case studies are presented. The first two cases discuss the relationship between implementation of RCM and the reduction of maintenance orders. The third case study demonstrates the reduction of maintenance cost, and the fourth case study demonstrates the increase in equipment availability with the implementation of RCM strategy.

2 RCM and Criticality Assessment

RCM is a systematic methodology for determining efficient and effective preventive maintenance tasks for systems in accordance with a specific set of procedures, and for determining maintenance schedule. Effective maintenance helps to achieve organization goals, and increase revenues by increasing equipment availability, performance and plant capacity, which leads to maximize the volume of productions and sales.

By definition, RCM is a structured process, methodology to determine the equipment maintenance plan required for any physical asset to ensure it continues to fulfill its intended present functions. Consequently, the objective of RCM is to determine the critical equipment in any system/process, and based on collected information and data; design a customized preventive/predictive maintenance plan for the organization. RCM initiatives, however, involve a great amount of resources, time, and energy. Thus the process is time-consuming and costly too [3–5].

Figure 1 shows a modified simple sketch of DuPont Stable Domain Model for maintenance strategy evolution. Historically the maintenance strategies developed. Initially, maintenance was only to repair the equipment once it failed. Then, corrective maintenance, and scheduled (preventive) maintenance have been introduced. Consequently, the predictive maintenance, which depends on equipment condition, has been adopted. Finally, the RCM is used. RCM is neither new methodology, nor contains new principles for performing maintenance. Actually, it is a structured approach of using the best methods, procedures, and disciplines. RCM controls the maintenance

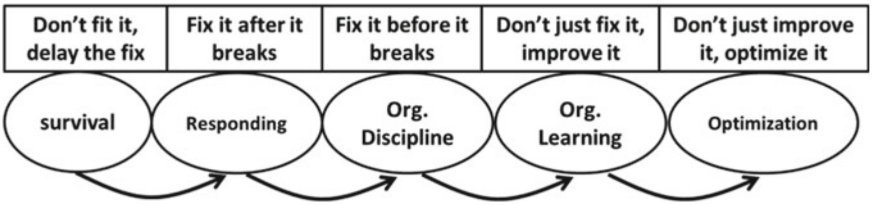
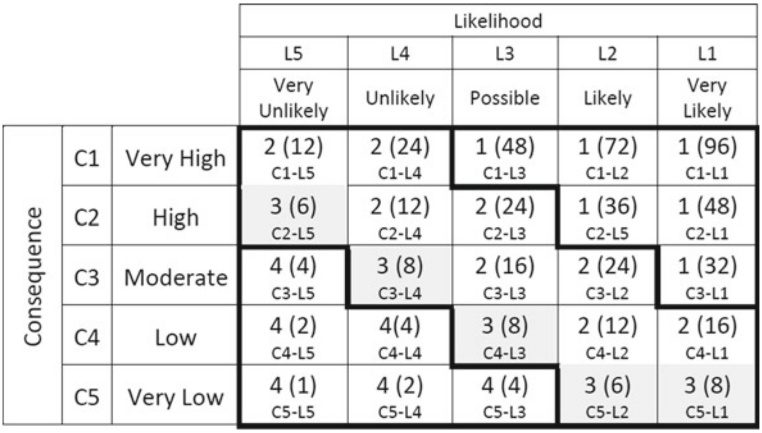


Fig. 1 Modified DuPont stable domain model for maintenance [6]

strategy at the level of plant or equipment type. The advantage of RCM is that it produces effective and robust planned maintenance programs, even in cases where there is little or no historical data. If RCM is properly applied, it can decrease the amount of scheduled maintenance work by a significant percentage. Proper introduction of RCM can lead to improved operating performance, greater safety, improved cost-effectiveness of maintenance, a comprehensive database, longer useful life of expensive capitals, better teamwork, and greater motivation of individuals.

Criticality analysis is an essential tool that provides valuable information for decision makers about work priority, and justifying resources for maintenance. It will ensure that resources are being spent in the most efficient way [1, 3, 7]. Sudden failure of major and critical systems/equipment can lead to extreme loss in production output. Therefore, it will be more cost-effective to predict the failure pattern of the critical systems/equipment so as to proactively plan, schedule and design maintenance tasks [8]. Plant, usually, composed of many types of equipment. A combination of equipment together forms a system. The systems in a plant have different levels of importance. So, it is imperative to decide which system to conduct t RCM study on. Deciding which system to study depends on different factors that will identify the criticality of each system. Usually, in a matured plant organization there is a clear guideline to define the criticality of systems using risk matrix (Fig. 2). A risk matrix is a rubric that is used during Risk Assessment to identify the different levels of risk as to the result of the probability categories and severity categories (L × C). This is a simple procedure to highlight the visibility of risks and assist management decision-making [9]. The following are the elements of a risk matrix.

(1) *Consequence Categories*: (a) Financial, (b) Reputation (c) Operational or Production Loss (d) Health and Safety and (e) Environmental;
(2) *Likelihood Scale*: It usually depends on the type of risk being frequency. Level 1 to Level 5 is shown in Fig. 2.



			Likelihood				
			L5	L4	L3	L2	L1
			Very Unlikely	Unlikely	Possible	Likely	Very Likely
Consequence	C1	Very High	2 (12) C1-L5	2 (24) C1-L4	1 (48) C1-L3	1 (72) C1-L2	1 (96) C1-L1
	C2	High	3 (6) C2-L5	2 (12) C2-L4	2 (24) C2-L3	1 (36) C2-L5	1 (48) C2-L1
	C3	Moderate	4 (4) C3-L5	3 (8) C3-L4	2 (16) C3-L3	2 (24) C3-L2	1 (32) C3-L1
	C4	Low	4 (2) C4-L5	4(4) C4-L4	3 (8) C4-L3	2 (12) C4-L2	2 (16) C4-L1
	C5	Very Low	4 (1) C5-L5	4 (2) C5-L4	4 (4) C5-L3	3 (6) C5-L2	3 (8) C5-L1

Fig. 2 Example of risk matrix

Table 1 Consequence and likelihood scoring

Consequence	Score	Likelihood	Score
C1 (Very high)	12	L1 (Very likely)	8
C2 (High)	6	L2 (Likely)	6
C3 (Moderate)	4	L3 (Possible)	4
C4 (Low)	2	L4 (Unlikely)	2
C5 (Very low)	1	L5 (Very unlikely)	1

Table 2 Level of risk

Measured risk score (L × C)	Risk level
32–96	RL 1 (Major)
12–24	RL 2 (Significant)
6–8	RL 3 (Minor)
1–4	RL 4 (Insignificant)

(3) *Consequence and Likelihood Scoring*: The scores of consequence describe the increase in the severity of the consequence (C1 to C5) with each level. Similarly the scores of the Likelihood describe the increase in the probability of likelihood (L) with each level as seen in Table 1.

The measured Risk (Likelihood × Consequence) score shall be fit in any one of the risk level ranges shown in Table 2.

3 Data Collection

SAP system was used to collect data required to perform the study. The numbers of different work orders and associated cost were collected. The objective is to find the impact of applying the RCM on those systems in terms of the number of corrective maintenance order (M1), number of breakdown maintenance order (M2) and number of preventive maintenance order (M3). Also, the associated cost with each of the raised orders whether it is increased or decreased. In order to perform the RCM analysis, the following seven basic attributes need to be considered by the RCM team [1, 10]: (1) Functions and desired performance of the (asset/system) (2) Functional Failures (3) Failure Modes (4) Failure Effect (5) Failure Consequences (6) Recommendations, and (7) Default Actions.

The first step in accomplishing the RCM study is the identification of the critical equipment. A team of technical staff assigned to carry out the Criticality Assessment. This team consists of experienced staff in operation, process, maintenance, mechanical, electrical and instruments, and external specialist if needed. Some specialists may be invited based on need such as Rotating specialist, inspection specialist, Sr. Reliability Engineer, etc. The team starts criticality assessment on system-by-system and equipment-by-equipment. Then, the team evaluates each system for the two

Table 3 Number of maintenance orders data for oxygen supply system

Year	M1 count	M2 count	M3 count	M1 cost	M2 cost	M3 cost
2007	4	0	0	3376.33	0	0
2008	8	0	15	18,769.15	0	25,071.88
2009	3	0	5	8133.05	0	10,074.10
2010	1	0	10	1267.04	0	16,419.56
2011	0	0	10	0	0	21,251.10
2012	2	0	13	5969.80	0	30,613.31
2013	1	0	16	673.54	0	36,791.94
2014	4	0	19	6889.46	0	37,940.52
2015	4	0	16	15,550.34	0	39,288.54

dimensions of risk matrix (consequence and likelihood). It is very important to know that the system consist of number of equipment and the criticality assessment will be done on the system as a whole including the related equipment [11, 12].

It is preferable that RCM study to be conducted for all equipment in the plant regardless of their criticality. There are number of equipment, however, in the plant that has a great impact once they fail. So, it is wise and logical to spend much of the time and effort on these critical systems or equipment and propose a suitable maintenance plan rather than spending much of time and effort in less critical equipment. A number of systems with risk level 1 (RL1) and risk level 2 (RL2) are selected in the case studies to demonstrate the effectiveness of applying RCM analysis, as they are the most critical failure type. The measured Risk (Likelihood × Consequence) the score should be fit in any one of the risk levels ranges shown in Table 2. Table 3 shows an example for Risk Matrix along with the Risk Levels that can be used to help in doing the criticality assessment.

4 Case Studies

The number of different maintenance orders were collected and analyzed. There are as follows:

a. corrective maintenance orders (M1)
b. breakdown maintenance orders (M2), and
c. planned maintenance orders (M3).

In addition, the associated costs of the different maintenance orders. In general, after implementing RCM team's recommendations, the number of corrective maintenance orders (M1) and the number of breakdown maintenance orders (M2) were reduced. On the other hand, the number of planned maintenance orders (M3) increased. The reduction in M1 and M2 orders had a great impact on optimizing cost and generation of M3 orders. Consequently, it helped in reducing failures, improve safety and environment, keeps the production continues, and sustains the organization's reputation. Also, it helped in improving equipment availability. Four case studies of implementing RCM are presented below.

A. *Case study No. 1: RCM and reduction of maintenance orders for Oxygen Supply System.*

This case study demonstrates the relationship between RCM implementation and the reduction in maintenance requests for the oxygen supply system. The daily production of this plant is 650 metric tons, which is equivalent to US Dollar 0.56 million per day. The function of this system is to supply oxygen to the Oxygen Mixer System (OMS) continuously for safe mixing of O_2 in cycle gas. One of the major equipment covered by this system is a relive valve. If the relive valve opens, this will cause the OMS to shut down which will lead to shut down the complete plant. This system has RL1 with a total risk score of 77.

Table 3 shows the number of maintenance orders performed on this system from 2007 to 2015 associated with the costs of maintenance orders. Figure 3 shows the trend of the three maintenance orders done on this system before and after the RCM study in 2012. It is very clear that the number of corrective maintenance (M1) is decreasing with time, which will increase the availability of the system as required. On the other hand, the number of preventive maintenance (M3) increased with time, which is a result of the RCM implementation. For the breakdown maintenance order (M2), no breakdown has been reported for this system since 2007. Although the number of M1 orders decreased with time a slight increase has been observed since 2011. Corrective maintenance cannot be eliminated totally. In this particular case, the increase in the number of M1 orders is due to implementing one of the RCM recommendations, which is "run to fail strategy". Such maintenance orders were not be registered as M3 orders because the date of the failure is not known and there was mitigation at the time of failure. Because it was not a planned order, it cannot be

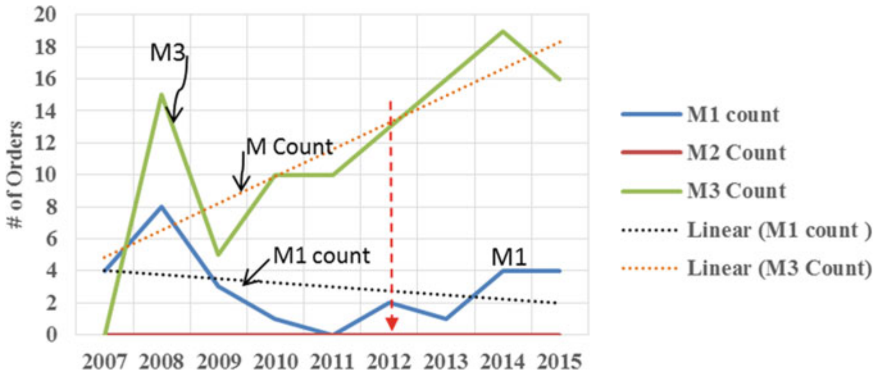


Fig. 3 Number of maintenance orders before and after RCM
]

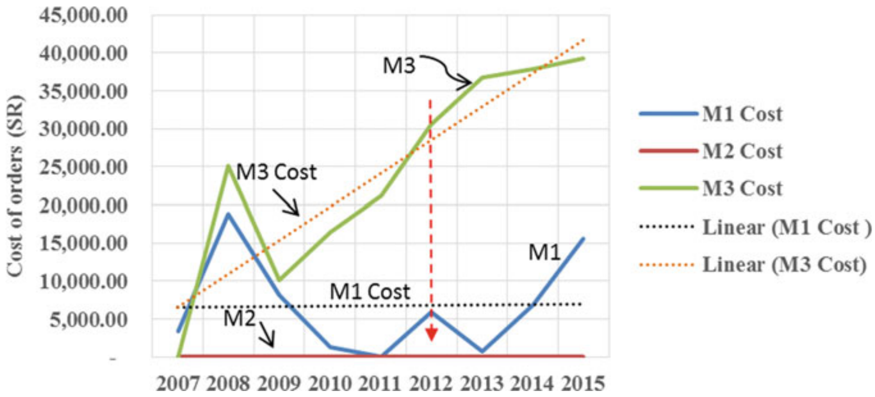


Fig. 4 Cost of maintenance order before and after RCM

registered as M3. In general, the trend of M1 orders is slopping downward whereas the M3 orders trend is sloping upward.

Figure 4 shows the costs related to these maintenance orders. It is very clear that the cost associated with the corrective maintenance orders (M1) is decreasing with time because the number of M1 orders have been decreased. As stated before, the associated cost started to increase after 2011, which is due to perform the M1 order of "run to fail strategy".

Such an increase in cost is justified for the organization's management. However, the associated cost with preventive maintenance orders is increasing with time because there are more M3 orders generated out of the RCM (Refer to Fig. 3). This increase in the cost of M3 orders are well justified, as this cost will have a direct impact on increasing the system's life and availability as required. No cost has been encountered for M2 orders, as there are no M2 orders happened. In general, the cost for the M1 orders is almost steady throughout the years whereas the cost associated with M3 orders is increasing with the years as seen by the trend line.

B. *Case study No. 2: RCM and reduction of Maintenance orders for Neutralization system.*

This system is used in Utility and Offsite plant. This plant is responsible to feed the other production plants with the utilities required for their production (cooling water, seawater, nitrogen, oxygen, steam …). The criticality assessment for this system is RL2 with a total risk score of 57. Table 4 shows the maintenance data related to this system. The RCM study implemented for this system in 2012 as highlighted in the table. Figure 3 shows the number of maintenance ordered performed on this system starting from 2007 till 2015.

It can be observed that the number of corrective maintenance orders (M1) is decreasing with time especially after the implementation of the RCM study in 2011. It is clearly shown that the number of preventive maintenance orders (M3) is having an increasing trend over the years. Figure 4 shows the trend of the cost over the years.

Table 4 Data for demineralized neutralization system

Year	M1 count	M2 count	M3 count	M1 cost	M2 cost	M3 cost
2007	25	0	28	29,720.11	0	12,312.53
2008	8	0	4	8437.85	0	5447.55
2009	26	0	1	19,796.42	0	2925.17
2010	33	0	8	41,048.25	0	9994.38
2011	25	1	29	73,422.50	4487.33	38,016.13
2012	11	0	29	39,679.59	0	43,027.33
2013	7	0	30	13,398.28	0	40,622.53
2014	4	1	53	9775.06	9726.00	65,451.81
2015	5	0	61	11,179.12	0	67,231.09

The reduction in the cost of the corrective maintenance orders (M1) is clear, which means that the equipment availability is in good condition. Also, the increasing trend of the preventive maintenance orders (M3) cost is well justified as that related to the decrease in the number of M1 orders and an increase in M3 orders (Figs. 5 and 6).

C. *Case study No. 3: RCM and maintenance cost.*

This case study demonstrates the potential saving generated from applying the RCM. Table 5 shows the potential economic benefits of applying RCM in 5 plants. The number of systems covered by RCM study is different from plant to plant. The first column, “Economic Risk before RCM” represents the risk of these systems if it fails as cost wise according to risk matrix. The systems were with risk level 1 (RL1: major) and risk level 2 (RL2: significant). After implementing the recommendations

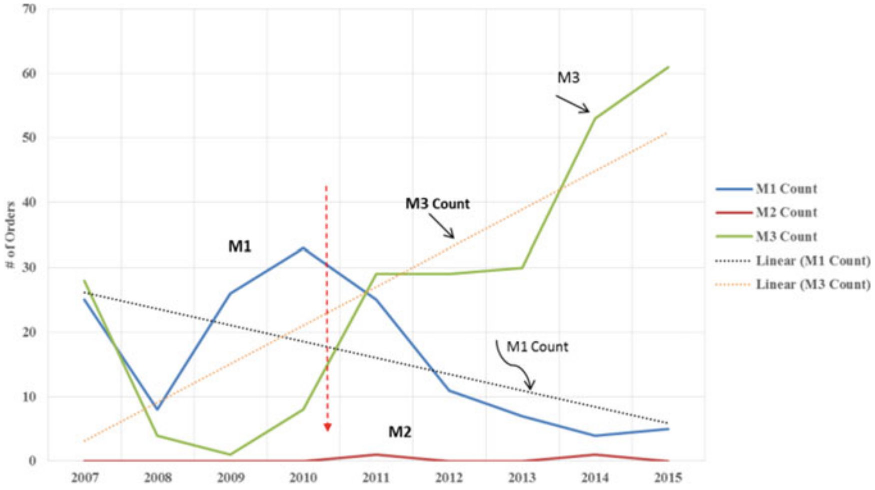


Fig. 5 Number of maintenance order before and after RCM

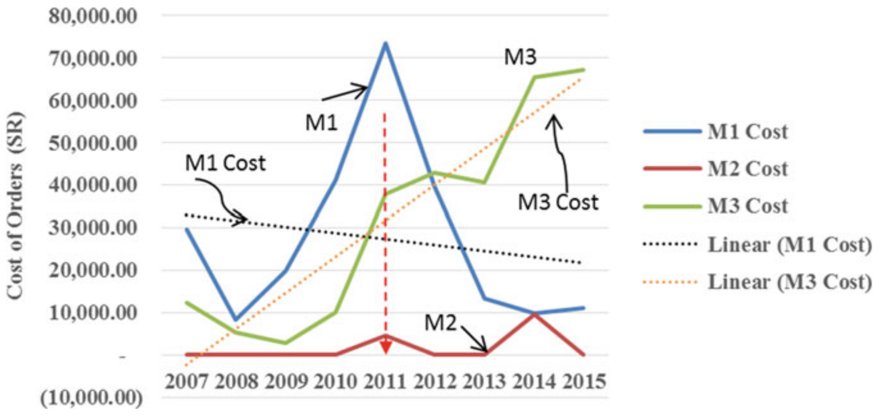


Fig. 6 Cost of maintenance order before and after RCM

Table 5 RCM potential economic benefits

Plant	Economic risk before RCM	Economic risk after RCM	Potential economic benefit	No. of systems
Plant 1	295,436,668.45	53,827,613.09	177,004,708.95	23
Plant 2	233,828,669.05	42,664,530.12	153,940,510.24	18
Plant 3	213,862,894.20	28,737,394.40	116,618,562.46	26
Plant 4	1,036,494,452.25	353,492,785.59	622,056,581.70	15
Plant 5	33,794,656.15	5,678,651.15	20,051,757.31	62
Total	1,813,417,340.10	484,400,974.34	1,089,672,120.65	144

of the RCM study, which might be (choosing the appropriate maintenance strategy, introducing PM, increasing PM, change frequency of PM, do regular inspection), the system showed significantly improvement.

This appears in the “Economic Risk after RCM” column, which means if the proposed recommendations have been implemented; the risk associated with this system is reduced to the noted value. The percentage of reduction in the economic risk due to implementing RCM study on these specific systems is seen. It is observed that there is a remarkable reduction in all plant figures. Figure 7 shows the potential economic benefits of implementing RCM recommendations for different plants. Although the number of systems is varying from plant to plant the associated potential economic benefits of RCM is not depending on how many systems covered by the RCM as shown in Table 5. Figure 8 shows the amount of reduction in the economic risk due to the implementation of the RCM recommendations by the year 2015. This result in saving the maintenance cost, improving the equipment’s reliability and the elimination of unwanted preventive maintenance. Some of the economic paybacks of RCM include: reduced production waste and maintenance cost: Increased reliability

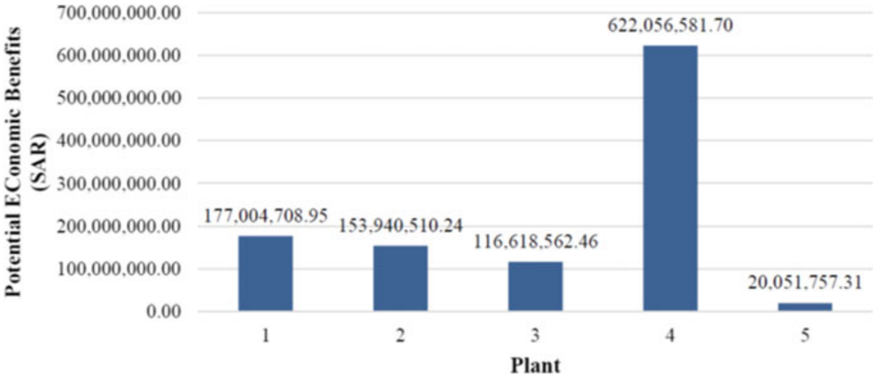


Fig. 7 RCM potential economic benefit (SAR) versus plants

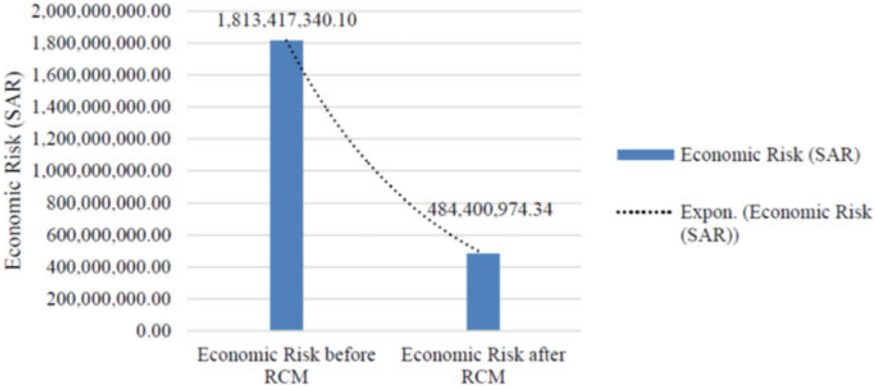


Fig. 8 Economic risk (SAR) before and after RCM

and availability of system/equipment c) decrees impact of failure; and Improved process and product quality.

It is important to assess whether the generated RCM recommendations have improved the equipment behavior or reduce the number of maintenance orders in the plants, which can be translated as the maintenance cost. Table 6 shows a case study conducted in one organization where the total maintenance cost has been collected from five different plants along with the cost of the RCM recommendations.

In general, the reduction can be also translated as a shift in moving the risk level of these systems from the risk level 1 (RL1: major) and risk level 2 (RL2: significant) to lower risk levels (RL3: minor or RL4: insignificant).

Figure 9 shows both the total maintenance cost and the cost of the RCM recommendations. It is clearly seen that the cost of implementing RCM recommendations has an increasing trend over the years. That will indicate that out of the total maintenance cost, RCM work orders have a quit good percentage of cost directed to

Table 6 Total maintenance cost to RCM recommendation cost

Year	Total maintenance cost	RCM recommendations cost	% of RCM cost contribution
2007	36,826,886.67	5,949,947.4	16.16
2008	41,383,500.86	6,480,760.76	15.66
2009	23,950,297.61	7,630,105.94	31.86
2010	39,517,968.07	12,505,842.33	31.65
2011	33,396,948.24	14,939,322.88	44.73
2012	37,826,446.56	15,334,293.74	40.54
2013	42,842,600.51	18,618,842.85	43.46
2014	43,401,531.99	19,348,677.75	44.58
2015	33,019,342.67	16,077,540.67	48.69
Total	332,165,523.2	116,885,334.3	35.19

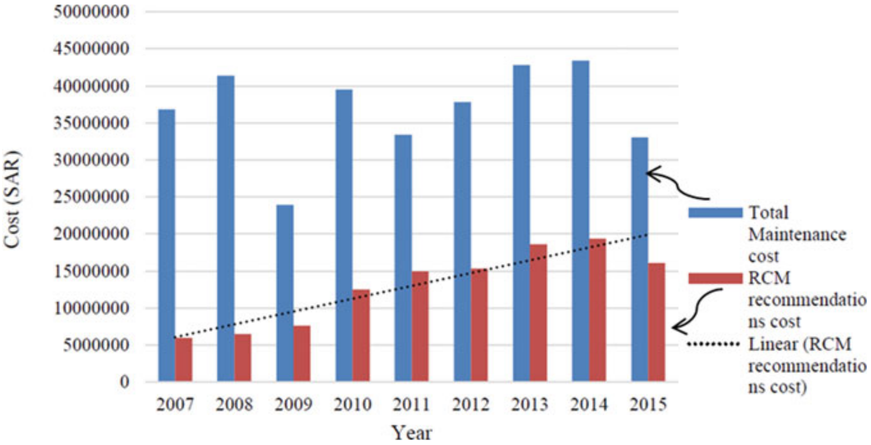


Fig. 9 Total maintenance cost to RCM recommendation Costs

the planned maintenance rather than the corrective maintenance and the breakdown maintenance. It also observed that the percentage of the RCM cost contribution to the total maintenance cost that the percentage of RCM recommendation cost is increasing year-after-year. The increase in the RCM recommendations cost compare to the total maintenance cost indicates the effectiveness of those recommendations in lowering the corrective maintenance orders (M1) and the breakdown maintenance orders (M2) and directing the resources to implement the RCM recommendations. These systems need some time to become more mature and the benefits will be realized more clearly. If the same manner of applying the RCM continues, the total maintenance cost will decrease and the cost of the RCM recommendations will increase with time.

Table 7 Equipment availability

Year	Sum of downtime (Hours)	Sum of equipment availability time (Hours)	Equipment availability (%)
2007	776.99	233,833,800	100.00
2013	17552.56	809,871,768	100.00
2014	17843.91	1,167,750,672	100.00

D. *Case study No. 4: RCM and equipment availability.*

This case study shows the improvement in the equipment availability ration that is considered a very important measure of the effectiveness of implementing the RCM recommendations. Availability is the time that a piece of equipment or system is capable of performing its intended functions divided by total time. It is usually expressed as a percentage. Table 7 shows the equipment availability data between the years 2007 and 2014. The sum of the equipment's downtime in hours and the total equipment availability in hours are shown. By applying the equipment availability concept, the equipment availability is shown in percentage. The downtime of the equipment cannot be eliminated totally; however, it can be reduced significantly by implementing RCM recommendations. Having standby equipment will reduce downtime and result in continuous operation with minimum interruption. Also, it will increase equipment reliability and improve equipment availability.

It is clearly observed that equipment availability between 2007 and 2014 showed an increasing trend. Figure 10 shows the equipment availability trend through the years having a continuous improvement, which will have a positive impact on the equipment to perform its intended function as required without any interruption. Ultimately, this will be a good support to have continuous production of the plant. Keep up the equipment available as required is one of the main objectives of the implementation of RCM recommendations.

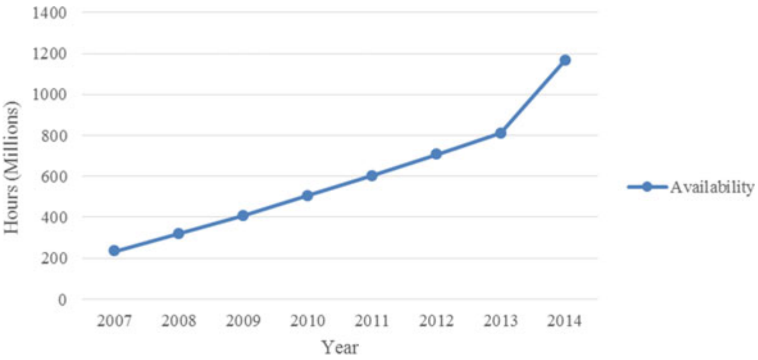


Fig. 10 Equipment availability

5 Conclusion

The results from the RCM analysis performed on several petrochemical plants showed a great improvement in operation and management satisfaction. As demonstrated in the case studies, there is a good reduction in the number of corrective maintenance orders (M1). This in turn leads to less number of failures, which results in increasing equipment and system availability. The results from the RCM analysis performed on several petrochemical plants showed a great improvement in operation and management satisfaction. As demonstrated in the case studies, there is a good reduction in the number of corrective maintenance orders (M1). This in turn leads to less number of failures, which results in increasing equipment and system availability. Also, it gives a clear indication that applying RCM reduces total maintenance cost. The number of preventive maintenance orders (M3) has increased for some systems after implementing the RCM recommendations. This is justifiable since the RCM team has recommended doing some preventive maintenance orders (inspection, overhaul, calibration …). Such actions lead to the improvement of equipment condition and service life and eventually in the reliability of the plants. Eliminating breakdown maintenance (M2) is a paramount goal of the maintenance team. Whenever there is M2 order, it means that plant is under shutdown, which is not preferred and not acceptable by the plant management. Although the cost associated with M2 orders appears to be very low or even negligible, the shown cost is only for repairing failed equipment and not representing the total cost due to M2 orders, which lead to shut down the complete plant. RCM teams will develop recommendations and time for implementation. Usually the time for implementation will be coupled with the planned shutdown. So, no process production is affected. The effect of implementing the RCM recommendations has been observed on the number of maintenance orders (corrective, breakdown, and preventive) and the cost associated with it. Also, the risk level of the critical systems has been declined after implementing the suggested RCM recommendations. Plant management needs to implement the RCM methodology on the critical systems as early as possible in order to decide and optimize the preventive maintenance strategies

Acknowledgements The authors appreciate the support of King Fahd University of Petroleum and Minerals during the study.

References

1. Reliability Centered Maintenance at: https://en.wikipedia.org/wiki/Reliability-centered_maintenance
2. J.T. Selvik, T. Aven, A framework for reliability and risk centered maintenance. Reliab. Eng. Syst. Saf. **96**(2), 324–331 (2011)
3. I.H. Afefy, Reliability-centered maintenance methodology and application: a case study. Engineering **2**, 863–873 (2010)

4. H.B. Jabar, Enhancing profitability through plant maintenance strategy, in 4th Reliability, Asset Management & Safety (RAMS) Conference 2008 "RAMS Implementation and Practice" 24–25 June 2008, Kuala Lumpur, Malaysia, pp. 1–12
5. R.G. Wilmeth, M.W. Usrey, Reliability-centered maintenance: a case study. Eng. Manag. J. **12**(4), 25–31 (2000)
6. The evolution of maintenance practices maintenance, http://www.lifetime-reliability.com/free-articles/maintenance-management/Evolution_of_Maintenance_Practices.pdf
7. SAE Standard JA1011_199908, Evaluation criteria for reliability-centered maintenance (RCM) processes, Issued 1999-08-01, Available at: http://standards.sae.org/ja1011_199908/
8. A.O. Kareem, Jewo, Development of a model for failure prediction on critical equipment in the petrochemical industry. Eng. Fail. Anal. **56**, 338–347 (2015)
9. Risk Matrix, Available at https://en.wikipedia.org/wiki/Risk_Matrix
10. S. Fore, A. Msipha, Preventive maintenance using reliability centered maintenance (RCM): a case study of a ferrochrome manufacturing company maintenance costs. Available at: http://www.idcon.com/resource-library/articles/reliability-vs-cost/548-reliability-improvements-drive.html
11. A.A. Bubshait, A. Basurrah, Impact of reliability centered maintenance on equipment availability and cost optimization, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, 3–5 July, 2019, London, UK, pp. 345–350
12. A.A. Bubshait, A. Al-Dosary, Application of lean six-sigma methodology to reduce the failure rate of valves at oil field, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014*, 22–24 October, 2014, San Francisco, USA, pp. 1010–1015

A Hybrid Methodology for Last Mile Delivery Strategy and Solution Selection at Smart Cities

Gülçin Büyüközkan and Deniz Uztürk

Abstract City logistics is a field, which can affect the residents' quality of life directly, refers to the optimization of logistics and transport activities while supporting the economic and social development of the city. "Last mile delivery" is the micro-level logistic operations in city logistics. It is directly related to the e-commerce activities of city residents, so; it continues to grow as a consequence of new connectivity solutions. This study focuses on two staged strategy and solution selection subject for the last mile delivery to augment its efficiency for companies and customers. The process is approached as a multi-criteria decision-making (MCDM) process. First, a SWOT analysis is conducted to obtain an in-depth evaluation of last mile delivery from a smart city perspective and generate strategies. Then the most suitable strategy for Istanbul city is selected by the 2-Tuple VIKOR approach. Plus, the SWOT factor prioritization is obtained by the Analytical Hierarchy Process (AHP). Finally, related to the SWOT analysis, suitable last mile delivery solution criteria are detected, and the suitable solution selection also implied with 2-Tuple VIKOR. The sensitivity analysis tests the plausibility of the suggested methodology, and the results are provided for this application.

Keywords 2-Tuple linguistic · Last mile delivery · Strategy selection · SWOT analysis · AHP · 2-Tuple VIKOR

G. Büyüközkan (✉)
Industrial Engineering Department, Galatasaray University, Çırağan Cad. no: 36, 34349 İstanbul, Turkey
e-mail: gbuyukozkan@gsu.edu.tr

D. Uztürk
Business Administration Department, Galatasaray University, Çırağan Cad. no: 36, 34349 İstanbul, Turkey
e-mail: duzturk@gsu.edu

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_17

1 Introduction

City logistics, being influenced by Industry 4.0, the internet of things (IoT), artificial intelligence, and wearable technologies, is trying to exist in the supply chain of the future. The constraints that directly affect customer satisfaction, such as delivery in a short time and delivery at the right time, have become more challenging.

Moreover, unexpected situations such as pandemics and severe natural events change the demand and the service level of the supply chain. City logistics are very vulnerable to these extreme changes; because its good flow is directly related to consumer preferences. At such times, technology and digitalization are the keys to achieve flexible and resilient processes in city logistics. Furthermore, the automation and robotics, which is emerged thanks to Industry 4.0, seem to be the future directions of city logistics.

In such circumstances, the last step in urban transport, which is the last mile delivery, has particular importance in city planning. Last mile delivery, is the final stage of the delivery process, from the delivery center or the factory to the end-user [1]. With the considerable growth of e-commerce in the market, the presence of last mile delivery in a supply chain becomes essential. The increase in technology also influences the last mile delivery unfavorably [2]. E-commerce is expanding due to the technology, and it challenges the last mile delivery with narrow timeframes and dense urban areas [2]. Furthermore, this urban density shows that a larger quantity of goods will be forecasted to be delivered shortly [3].To be able to handle this expansion in good-flow, innovative solutions such as collaborative urban logistics, optimization in routing, proximity stations, and innovative vehicles are generated to optimize last mile delivery, which is the least profitable stage of the supply chain [3].

Motivated by these innovative solution proposals and the challenging situation of last mile delivery, this study focuses on analyzing the last mile delivery from a smart city perspective. By doing that, this chapter aims to create an easy-to-apply guideline for companies or municipalities to generate suitable last mile solutions in the urban areas.

This study wanted to assess the last mile delivery subject around the notion "smart city" because, the cities in the future are expected to use and adapt to developing technologies. Besides, when today's situation with COVID-19 is analyzed, the need for smart solutions in the city are crucial [4]. The technological adaptation of cities will be the significant targeted area for municipalities in the future. Consequently in this chapter, the last mile delivery subject is evaluated from a smart city perspective.

In the literature, still, it is not clear that last mile delivery covers business-to-business (B2B) deliveries to small businesses [5]; so, this study only focuses on business-to-customer (B2C) deliveries in cities as a result of its upward trend [6]. In this study, the strategy generation is handled by SWOT analysis, and the prioritization of SWOT factors is calculated by the Analytical Hierarchy Process (AHP). As the second stage, the appropriate solution for the suited strategy is selected by the 2-Tuple linguistic integrated VIKOR methodology.

VIKOR technique is provided with its 2-Tuple linguistic extensions to empower their ability to deal with linguistic data. Also, the 2-Tuple linguistic model offers a flexible linguistic decision-making environment to the decision-makers (DMs). The 2-Tuple approach gives the possibility to assign different granulated linguistic term sets to DMs according to their experience/knowledge about the subject.

The remainder of this chapter is organized as follows: Sect. 2 provides a literature review about the last mile delivery studies conducted in recent years. Section 3 gives the methodology with SWOT analysis for last mile delivery from the perspective of smart cities, and it also provides the generated strategies. Section 4 presents the application of the provided methodology for strategy and solution selection. Finally, Sect. 5 and Sect. 6 give managerial implications and conclusions, respectively.

2 Last-Mile Delivery

The last mile delivery subject is prevalent in literature for the last few years. When the "last mile delivery" word is searched on the Scopus database for the last five years, three hundred and four different works have been focused on the last mile delivery as a subject. Most of the pieces have been published between the years 2017 and 2019. However, every work has approached the same subject from different perspectives and have suggested various solutions for city freight transportation with last mile delivery. Between them, only thirty-eight published articles have focused last mile delivery from a smart perspective. In addition, thirteen of them were related last mile delivery with smart cities. Table 1 gives the distribution of these thirteen articles according to the fundamental topics under the last mile delivery subject. These articles are mostly published in 2018, 2019, and 2020 related to technological developments. In this study, the main aim is to focus on the last mile delivery subject from the smart city perspective to be able to create intelligent solutions.

The concept of "smart city" is emerged from the integration of technology with the city. The definition of a smart city is done in many different ways. As a concept, it is a bit fuzzy and comes in different ways with varying features in the literature [19].

Table 1 Recent studies about last mile delivery with smart cities

Topic	References
Network/route optimization	[7–9]
Sustainable last mile delivery	[10–14]
Parcel deliveries	[15, 16]
Electric vehicles	[17]
Collaborative last mile delivery	[18]
Last mile delivery innovations	[3]

The adjective smart in urban planning is generally addressed in a strategic dimension [19]. It refers to the adoption of different ideologies, strategies, and approaches in the planning of the city and uses markup styles. When it is addressed in technological wisdom, it refers to the technological integration of whole sub-systems of the city such as transportation, infrastructure, buildings, etc.

In the literature, it exists three studies that assessed the last mile delivery subject with smart cities. Two of them published as conference papers in 2015. In the first one, Navarro et al. have proposed new models for smart city logistics and applied them for Spain [20]. In the second one, Lindawati et al. suggested a platform with minimum data sharing based on a collaboration for city logistics [21]. As a third study, in 2017, Dispenze et al. suggested an infrastructure study for sustainable mobility in cities [22].

Amid different studies investigated last mile delivery with smart cities, there is a lack of bottom-up review of the last mile delivery subject from a smart city perspective in order to create smart strategies. It is clear that today, in this digitalization age, cities are getting more intelligent and interactive places where all systems are connected. At this point, last mile delivery, which is the critical part of the city logistics, must be well evaluated with smart city notion to be able to better adapt to the future's smart solutions.

3 Methodology

In this chapter, the last mile delivery subject is investigated with a smart city notion. In the first step, to be able to have an in-depth assessment of the issue, a SWOT Analysis has been conducted. This analysis has provided an environment for a better understanding of the subject and generate the selection criteria for the smart solution selection.

3.1 SWOT Analysis

The SWOT analysis, which is widely preferred by business planners, is a tool for evaluating and measuring the internal and external environment of the company. SWOT consists of the initials of the words: strength, weakness, opportunities, and threats [23]. The SWOT analysis has become a widely used method for an in-depth understanding of many subjects, not only on companies but also to see the interrelations of both the concept and the external systems in detail. The components of the SWOT analysis are described as follows [23]. We try to answer these questions for a better understanding of the SWOT components:

Strengths: What are the advantages? What can be best done about this? What resources and contacts can we reach?

Weaknesses: What are the missing parts? What can be improved more?

Opportunities: What opportunities can be found in the sector? What are the trends that can create new opportunities?

Threats: What are the obstacles that can be encountered in the external environment? Do these barriers affect the sustainability of the system?

3.2 *SWOT Analysis for Last Mile Delivery at Smart Cities*

With the result of the literature review and the help received from the experts, the last mile delivery from the smart city perspective was examined comprehensively. The obtained results of the SWOT Analysis are presented in Table 2.

In this study, based on the data obtained from the SWOT analysis, it is aimed to make a suitable last mile delivery strategy selection following an appropriate solution selection, with a 2-Tuple integrated VIKOR method.

Strengths indicate the advantages achieved when the last mile delivery in smart cities is implemented. Particularly in the last mile delivery, the benefits gained by the companies which are looking for intelligent solutions were mentioned. These benefits, as discussed above, have essential features in terms of financial and customer satisfaction.

In the Weaknesses, it is considered as the problems that may arise on the firm side when the subject is applied. Here, the importance of technological infrastructure and trained employees constitute a critical point.

Table 2 SWOT analysis for last mile delivery from smart city perspective

	Strengths	Weaknesses
Internal factors	S1: Less cost for delivery companies [23] S2: Increased rate of on-time deliveries [23] S3: Detection of the exact location of the cargo in the transport chain [23] S4: Energy-saving and CO_2 emission reduction with less truck transportation [23]	W1: Confidentiality and security problem due to information sharing [23] W2: The system does not function correctly unless it has a robust technological infrastructure [23] W3: The need for staff to be well trained in new technologies and complex systems [23] W4: High investment cost [24]
External factors	Opportunities	Threats
	O1: Strengthening stakeholder cooperation [24] O2: The city has a positive power in the fight against climate change [24] O3: A city with a high level of prosperity with less traffic [24] O4: Help reduce the crime rate in the city [24]	T1: Weak stakeholder cooperation [3] T2: Legal and political barriers to transportation [3] T3: Poor signal-induced accidents [3]

Table 3 Last mile delivery strategies concerning smart city

A#	Strategies
A1	Creating a fleet with more efficient delivery and low carbon emission
A2	Investing in a robust IT infrastructure
A3	Investing in a blockchain based system for data security
A4	Training staff about the sustainability
A5	Making R&D studies or searching for a technologic partner
A6	Using a collaborative platform
A7	Build a team to analyze the changing customer expectations
A8	Working with start-ups to generate efficient technologies with low investment costs

Opportunities focus on the gains of the city and the city's stakeholders, as the evaluation was made from the smart city perspective. This section focuses on the city's profits more than the companies. The most prominent criterion in the Opportunities section is the increase in the quality of life of the city and the increase in the level of welfare.

When the Threats section is evaluated, the challenges that the smart city may face in the last mile delivery practices are discussed. The last step is to indicate the problems that may be encountered for delivery. Here, the weak stakeholder collaboration in the city is one of the major threats identified for the final step.

3.3 Strategies for Last Mile Delivery for Smart Cities

The SWOT analysis conducted in the previous section gave the possibility to generate related strategies for last mile delivery. Accordingly, Table 3 presents the list of the determined strategies.

In the future sections, these strategies and the SWOT analysis will be a helpful guide to generate last mile delivery solution selection criteria. Strategies A1 and A2 are detected for the intersection of Opportunities and Strengths factors. A3 and A4 are for the intersection of Opportunities and Weaknesses. A5 and A6 are for Threats and Strengths and finally, A7 and A8 are for Threats and Weaknesses.

The AHP method is a widely used technique in decision making. It is firstly introduced by Saaty in 1982 [25]. Il also usually used with group decision making as in this chapter. In this method, the fuzziness of the decision making is handled by assigning the linguistic terms to numbers between 1-9. By this assignment, the hierarchical complexity and the vagueness are buried in the decision making. The details of the AHP steps are detailed in [25].

VIKOR, the Serbian name; Vlse Kriterijumska Optimizacija I Kompromisno Resenje, is an MCDM approach to select the most suitable alternative among multiple alternatives [26]. Its selection process is based on the compromise solution, which

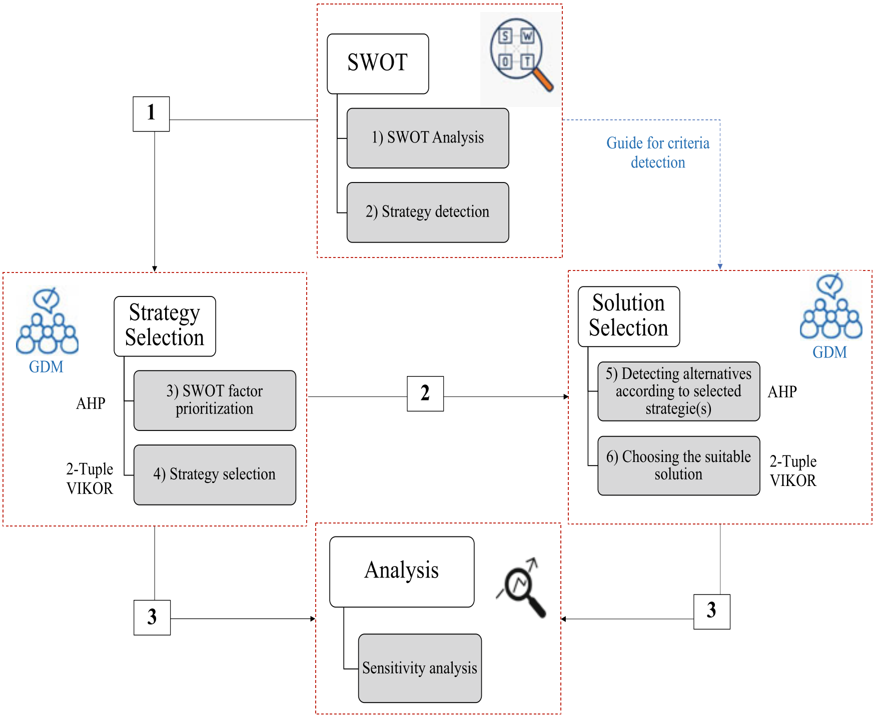


Fig. 1 Proposed framework

is firstly handled by Yu in 1973 [27]. In this technique, the compromise solution is the closest solution to the ideal solution [28]. In this chapter, the VIKOR method is suggested with its 2-Tuple extension. Due to the page limitations the details of the 2-Tuple VIKOR technique steps can be reached from [26]. The detailed process of the framework are provided in Fig. 1.

4 Case Study: Last Mile Delivery Strategy and Solution Selection for Istanbul

Nowadays, we know that cities are in a transformation to provide the right environment for their inhabitants. In Istanbul, it is the same, and it aims to provide a better environment for its inhabitants with plans such as the Climate Change Action Plan and Air Quality Action Plan.

An application of the proposed framework was conducted for the city of Istanbul by following the steps given in the previous section.

Steps 1–2: A literature research and a SWOT Analysis have been conducted as in the previous sections. Accordingly, the strategies are generated as in Sect. 3.3.

Finally, a decision-making group is formed from three experts, which are related to logistics, technologies, and sustainability issues. The constructed SWOT structure is given in Fig. 2.

Step 3: SWOT factor prioritization is obtained by applying AHP for the Strengths, Weaknesses, Threats, and Opportunities dimensions and their sub-dimensions. Table 4 gives an evaluation of the first DM as an example. It also gives the ultimate weights of the dimensions. Also, Fig. 3 shows the weighting of the criteria.

Step 4: This step is strategy selection. The 2-Tuple integrated VIKOR [30] is applied to obtain a suitable strategy for last mile delivery in smart cities. A linguistic set is defined as the decision-making group to express their opinions about the subject.

They have given their assessments in a five-labeled linguistic set which consists of: *No influence (N)-Low influence (L)-Medium influence (M)-High influence (H)-Very High influence (VH).*

Here Table 5 presents the ultimate aggregated weighted decision matrix. Furthermore, Fig. 4 gives the S, R, and Q ranking of alternatives. Accordingly, when the VIKOR conditions are considered, the most suitable strategy is A1; however, A1, A3, A4, A6, A7 form a feasible solution group.

Step 5 and 6: As a result of the strategy selection, we obtained A1 and A6 as the appropriate ones. They suggest investing in efficient fleets for deliveries and creating collaborative platform for market to generate regulations. Accordingly, three different alternatives have been detected for the practical solution to generate these strategies [29]. And the evaluation criteria are generated from the SWOT analysis. The criteria are: *Decreasing cost (C1), Increasing customer satisfaction (C2), Improved traceability (C3), More adaption to climate change (C4), augmented information security (C5), Issued related to lack of infrastructure (C6), Well-trained staff (C7), High cost of investment (C8), Barriers to the operation of the system (C9), Powerful cooperation with stakeholders (C10), High level of welfare in the city (C11), Gain of efficiency (C12), Ease of multiplication (C13), Connectivity (C14)* [30]. The same linguistic set is defined to the DM group.

First, fixed or temporary smart cabinets (Solution 1) are identified as an alternative [29]. This alternative includes temporary or permanent lockers, which are placed in the busiest part of the city and can be reached by public transport. These cabinets are connected to a cloud system and are used with an application on the users' phone.

When the delivery is placed in the smart closet, a notification is sent to the customer via the system, and the customer can access the delivery at any time of the day with the code sent. Another alternative is autonomous electric vehicles (Solution 2). These vehicles are usually small and traffic-independent vehicles [31]. They have a system that manages the road information and the locations to be traveled, and they have the least impact on the environment as it works with electricity. Figure 5 gives the S, R, Q ranking for solution alternatives.

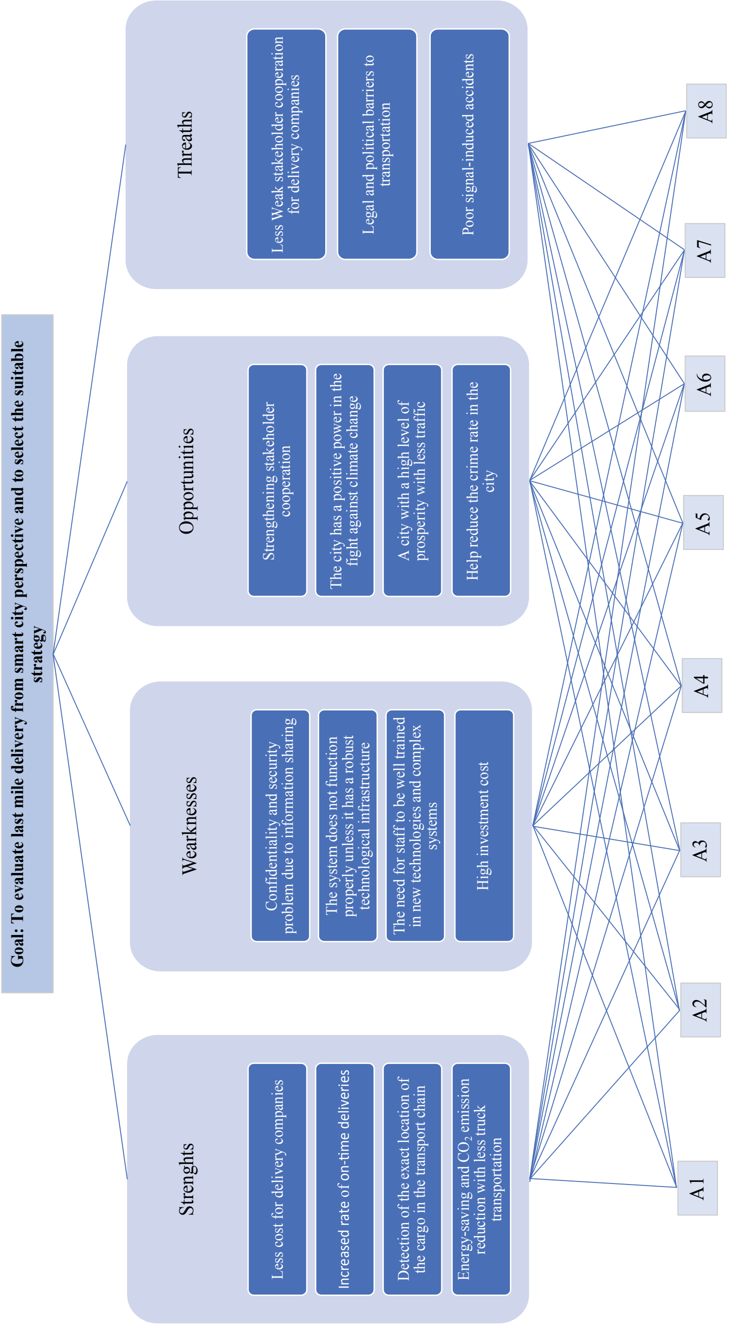

Goal: To evaluate last mile delivery from smart city perspective and to select the suitable strategy
Strenghts
Less cost for delivery companies
Increased rate of on-time deliveries
Detection of the exact location of the cargo in the transport chain
Energy-saving and CO_2 emission reduction with less truck transportation
Wearknesses
Confidentiality and security problem due to information sharing
The system does not function properly unless it has a robust technological infrastructure
The need for staff to be well trained in new technologies and complex systems
High investment cost
Opportunities
Strengthening stakeholder cooperation
The city has a positive power in the fight against climate change
A city with a high level of prosperity with less traffic
Help reduce the crime rate in the city
Threaths
Less Weak stakeholder cooperation for delivery companies
Legal and political barriers to transportation
Poor signal-induced accidents
A1
A2
A3
A4
A5
A6
A7
A8

Fig. 2 Constructed hierarchical SWOT model for case study

Table 4 Evaluation of the first DM for AHP

	Strength	Weakness	Opportunity	Threat	Weights
Strength	1.000	0.143	0.500	0.125	0.06
Weakness	7.000	1.000	1.000	0.500	0.29
Opportunity	2.000	1.00	1.000	0.500	0.22
Threat	4.000	2.000	2.000	1.000	0.43

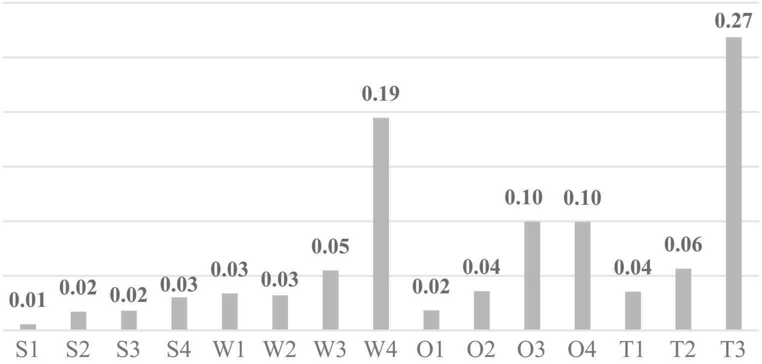


Fig. 3 Criteria weights

5 Managerial Implications

In this chapter, an MCDM approach is applied to evaluate and select a suitable strategy for last mile delivery. Since the suggested methodology is an expert-based system, it relies on the knowledge and the experience of the DM on the subject. In the case study, the last mile delivery SWOT analysis is implied from the smart city point of view for Istanbul. Five different strategies are selected in the feasible set. A1, which is the "Creating a fleet with small and low carbon emission vehicles for parcel deliveries," strategy seems to be more suitable since it is ranked as the first alternative in S and Q scores.

The strategy ranking may vary from city to city, depending on their digitalization level and culture. However, the sensitivity analysis is generated to test the robustness and the plausibility of the provided methodology in the following section.

A sensitivity analysis is conducted for strategy and suitable solution selection to test the plausibility of the suggested methodology. Due to the page limitations, only the results of the strategy selection are presented in Fig. 6; since it has a direct effect on solution selection. For the analysis, three different cases are generated in which the weighting of the main dimensions (strengths, weaknesses, opportunities, threats) change.

The first dark graph presents the results of the applied case study where the weights of the dimensions are obtained approximately $(0.1, 0.3, 0.2, 0.5)$ for strengths, weaknesses, opportunities, threats, respectively. The other three cases of weightings are as follows:

Table 5 Ultimate aggregated evaluation matrix for strategy selection

a

	S1	S2	S3	S4	W1	W2	W3	W4
A1	(S_1, −0.17)	(S_0, 0.33)	(S_1, 0)	(S_0, 0)	(S_0, 0)	(S_0, 0)	(S_1, 0.17)	(S_1, −0.13)
A2	(S_1, 0)	(S_1, 0)	(S_0, 0)	(S_0, 0.50)	(S_0, 0.43)	(S_0, 0.43)	(S_0, 0)	(S_1, −0.13)
A3	(S_0, 0.33)	(S_1, 0)	(S_1, −0.13)	(S_1, 0)	(S_1, 0)	(S_1, 0)	(S_0, 0)	(S_0, 0)
A4	(S_1, −0.13)	(S_0, 0)	(S_1, 0)	(S_1, 0)	(S_0, 0.14)	(S_0, 0.29)	(S_1, 0)	(S_0, 0.33)
A5	(S_0, 0)	(S_1, 0)	(S_1, 0)	(S_1, 0)	(S_1, −0.29	(S_0, 0.43)	(S_0, 0)	(S_1, −0.13)
A6	(S_1, −0.13)	(S_1, −0.13)	(S_0, 0.33)	(S_0, 0.50)	(S_1, −0.29)	(S_0, 0.14)	(S_0, 0)	(S_0, 0.33)
A7	(S_1, −0.23)	(S_1, 0)	(S_1, −0.23)	(S_1, 0)	(S_1, 0)	(S_0, 0.43)	(S_0, 0.33)	(S_1, 0.17)
A8	(S_0, 0.33)	(S_0, 0.33)	(S_1, −0.13)	(S_0, 0.50)	(S_0, 0.43)	(S_0, 0.43)	(S_0, 0.33)	(S_1, 0)

b

	O1	O2	O3	O4	T1	T2	T3
A1	(S_1, 0)	(S_0, 0)	(S_0, 0)	(S_1, 0)	(S_0, 0)	(S_0, 0)	(S_0, 0)
A2	(S_0, 0)	(S_1, 0)	(S_0, 0.33)	(S_1, 0)	(S_0, 0.14)	(S_0, 0.25)	(S_0, 0.14)
A3	(S_0, 0)	(S_0, 0.43)	(S_1, −0.13)	(S_0, 0)	(S_0, 0.43)	(S_0, 0.38)	(S_0, 0.43)
A4	(S_1, −0.13)	(S_0, 0.14)	(S_0, 0.33)	(S_1, 0)	(S_0, 0.29)	(S_0, 0.25)	(S_0, 0.14)
A5	(S_0, 0.33)	(S_1, −0.29)	(S_1, 0)	(S_1, 0)	(S_0, 0)	(S_0, 0)	(S_1, 0)
A6	(S_0, 0)	(S_0, 0.14)	(S_1, −0.23)	(S_0, 0)	(S_1, 0)	(S_1, 0)	(S_0, 0.43)
A7	(S_0, 0.50)	(S_0, 0.43)	(S_1, −0.17)	(S_0, 0)	(S_1, 0)	(S_0, 0.50)	(S_0, 0.43)
A8	(S_0, 0.33)	(S_0, 0.43)	(S_1, 0)	(S_1, 0)	(S_0, 0)	(S_0, 0)	(S_1, 0)

Case 1: (0.5, 0.1, 0.2, 0.3)-Case 2: (0.3, 0.5, 0.1, 0.2)-Case 3: (0.2, 0.3, 0.5, 0.1).

As mentioned in Fig. 6, the ranking of the strategies varies according to the weights of the dimensions. However, the feasible solution set is not differing a lot. In most cases, A1 and A6 are still in the feasible solution set. Also, our selected solution is still valid for these cases. Case 1 is the one with a more diversified ranking. In this case, the highest weight is given to the opportunities dimension, but still, there is no significant deviation from the real case study.

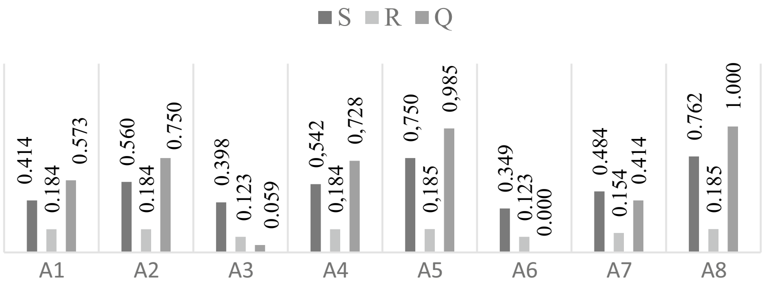


Fig. 4 *S, R, Q* values for strategies

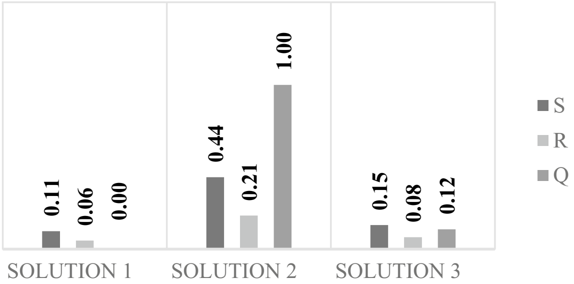


Fig. 5 *S, R. Q* values for solution alternatives

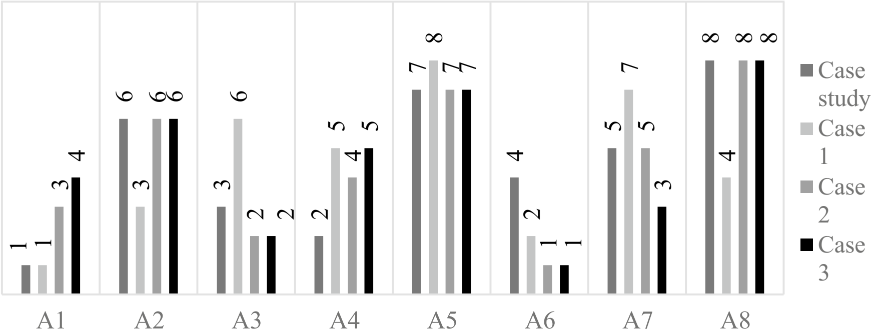


Fig. 6 Sensitivity analysis

6 Conclusions

Smart solutions for cities are topical subjects due to emerging technologies. Moreover, today we are aware of possible severe risks such as pandemics and natural disasters for humanity. And we experienced that the smart solutions are the essential exit to adapt and mitigate against these risks.

Transportation, which is often a prominent hot spot for most cities, is on the mend as a consequence of new technologies. Emerging optimization techniques such as smart routing, congestion planning help to reduce the negative impacts of city transportation for the residents.

Last mile delivery is one the critical part of the city transportation, which constitutes of commercial operations. It has become indispensable for logistics companies and residents due to active e-commerce activities. Motivated by this growth of last-mile logistics for B2C applications, this study is proposed a strategy selection methodology. In this methodology, a hybrid process of AHP and 2-tuple VIKOR is implied. Accordingly, investing in low carbon emission vehicles is selected as the most suitable strategy for the case. Also, the possible solution selection is indicated to generate a smart solution for last mile deliveries. In this case, the alternative one, which is the temporary smart cabinets, is selected as the most suitable smart solution. This alternative is also valuable to eliminate congestion in urban areas and also provides more customer satisfaction without time limitations.

The fact that the new generation of urban logistics has an approach that unites most stakeholders and directs them to cooperate with each other, it may open the horizons to the new research areas about the distribution of the tasks of the stakeholders and how cities can easily implement this system on their own.

Acknowledgements This work has been supported by the Scientific Research Projects Commission of Galatasaray University under grant number 19.102.002. We gratefully thank our experts.

References

1. Cerasis: The Ultimate Guide to Last Mile & White Glove Logistics. https://cerasis.com/2017/10/05/e-book-last-mile-and-white-glove-logistics/. Last accessed 21 Sept 2018
2. CBRE, *Last Mile\City Logistics* (2017)
3. L. Ranieri, S. Digiesi, B. Silvestri, M. Roccotelli, A review of last mile logistics innovations in an externalities cost reduction vision. Sustainability **10**, 782 (2018). https://doi.org/10.3390/su10030782
4. Z. Allam, D.S. Jones, On the coronavirus (COVID-19) outbreak and the smart city network: universal data sharing standards coupled with artificial intelligence (AI) to benefit urban health monitoring and management. Healthcare **8**, 46 (2020). https://doi.org/10.3390/healthcare8010046
5. I. Cardenas, Y. Borbon-Galvez, T. Verlinden, E. Van de Voorde, T. Vanelslander, W. Dewulf, City logistics, urban goods distribution and last mile delivery and collection. Competition Regul. Netw. Indu. **18**, 22–43 (2017). https://doi.org/10.1177/1783591717736505
6. Statistics About Retail Marketing and Consumer Shopping Trends V12. https://www.v12data.com/blog/50-statistics-about-retail-marketing-and-consumer-shopping-trends/. Last accessed 12 March 2019
7. F.M. Bergmann, S.M. Wagner, M. Winkenbach, Integrating first-mile pickup and last-mile delivery on shared vehicle routes for efficient urban e-commerce distribution. Transp. Res. Part B: Methodological **131**, 26–62 (2020). https://doi.org/10.1016/j.trb.2019.09.013

8. J.J.Q. Yu, W. Yu, J. Gu, Online vehicle routing with neural combinatorial optimization and deep reinforcement learning. IEEE Trans. Intell. Transp. Syst. **20**, 3806–3817 (2019). https://doi.org/10.1109/TITS.2019.2909109
9. Y. Wang, D. Zhang, Q. Liu, F. Shen, L.H. Lee, Towards enhancing the last-mile delivery: An effective crowd-tasking model with scalable solutions. Transp. Res. Part E Logistics Transp. Rev. **93**, 279–293 (2016). https://doi.org/10.1016/j.tre.2016.06.002
10. R. de Kervenoael, A. Schwob, C. Chandra, E-retailers and the engagement of delivery workers in urban last-mile delivery for sustainable logistics value creation: leveraging legitimate concerns under time-based marketing promise. J. Retail. Consum. Serv. **54**, 102016 (2020). https://doi.org/10.1016/j.jretconser.2019.102016
11. A. Giret, C. Carrascosa, V. Julian, M. Rebollo, V. Botti, A crowdsourcing approach for sustainable last mile delivery. Sustainability (Switzerland) **10**, 4563 (2018). https://doi.org/10.3390/su10124563
12. L.G. Marujo, G. Goes, M.A. D'Agosto, A.F. Ferreira, M. Winkenbach, R.A.M. Bandeira, Assessing the sustainability of mobile depots: the case of urban freight distribution in Rio de Janeiro. Transp. Res. Part D-Transp. Environ. **62**, 256–267 (2018). https://doi.org/10.1016/j.trd.2018.02.022
13. G.V. Goes, D.N. Schmitz, R.A.D.M. Bandeira, C.M. De Oliveira, M.D.A. D'Agosto, Sustainability in the last mile of urban freight transport: The role of vehicle energy efficiency. Sustentabilidade em Debate. **9**, 134–144 (2018). https://doi.org/10.18472/SustDeb.v9n2.2018.27418
14. C.M. de Oliveira, R.A. De Mello Bandeira, G.V. Goes, D.N. Schmitz Goncalves, M.D.A. D'Agosto, Sustainable vehicles-based alternatives in last mile distribution of urban freight transport: a systematic literature review. Sustainability **9**, 1324 (2017). https://doi.org/10.3390/su9081324
15. T.B.T. Nguyễn, T. Bektaş, T.J. Cherrett, F.N. McLeod, J. Allen, O. Bates, M. Piotrowska, M. Piecyk, A. Friday, S. Wise, Optimising parcel deliveries in London using dual-mode routing. J. Oper. Res. Soc. **70**, 998–1010 (2019). https://doi.org/10.1080/01605682.2018.1480906
16. F. Wang, F. Wang, X. Ma, J. Liu, Demystifying the crowd intelligence in last mile parcel delivery for smart cities. IEEE Netw. **33**, 23–29 (2019). https://doi.org/10.1109/MNET.2019.1800228
17. R.A. de Mello Bandeira, G.V. Goes, D.N. Gonçalves, D.A. Márcio de Almeida, C.M. de Oliveira, Electric vehicles in the last mile of urban freight transportation: a sustainability assessment of postal deliveries in Rio de Janeiro-Brazil. Transp. Res. Part D Transport Environ. **67**, 491–502 (2019). https://doi.org/10.1016/j.trd.2018.12.017
18. A. Bhasker, S.P. Sarmah, T. Kim, Collaborative last-mile delivery and pick-up in city logistics. Int. J. Logistics Syst. Manage. **34**, 533–553 (2019). https://doi.org/10.1504/IJLSM.2019.103518
19. T. Nam, T.A. Pardo, Conceptualizing smart city with dimensions of technology, people, and institutions, in *Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in Challenging Times—dg.o '11* (ACM Press, College Park, Maryland, 2011), p. 282. https://doi.org/10.1145/2037556.2037602
20. C. Navarro, M. Roca-Riu, S. Furio, M. Estrada, Designing new models for energy efficiency in urban freight transport for smart cities and its application to the Spanish case, in *Ninth International Conference on City Logistics*, ed. by E. Taniguchi, R.G. Thompson (Elsevier Science Bv, Amsterdam, 2016), pp. 314–324
21. Lindawati, C. Wang, W. Cui, N. Hari, *Feasibility Analysis on Collaborative Platform for Delivery Fulfillment in Smart City* (IEEE, New York, 2015)
22. G. Dispenza, V. Antonucci, F. Sergi, G. Napoli, L. Andaloro, Development of a multi-purpose infrastructure for sustainable mobility. A case study in a smart cities application, in *Leveraging Energy Technologies and Policy Options for Low Carbon Cities*, ed. by J. Yan, S.K. Chou, H. Li, V. Nian (Elsevier Science Bv, Amsterdam, 2017), pp. 39–46
23. K. Papoutsis, M. Gogas, E. Nathanail, *Urban Distribution Concepts: A SWOT Analysis on Best Practices of Urban Logistics Solutions* (2012), p. 19

24. Deloitte: Annual Industry Report, https://www.mhi.org/publications/report. Last accessed 21 Sept 2018
25. T.L. Saaty, The analytic hierarchy process: a new approach to deal with fuzziness in architecture. Architectural Sci. Rev. **25**, 64–69 (1982). https://doi.org/10.1080/00038628.1982.9696499
26. S. Opricovic, G.-H. Tzeng, Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. Eur. J. Oper. Res. **156**, 445–455 (2004). https://doi.org/10.1016/S0377-2217(03)00020-1
27. P. Yu, A class of solutions for group decision problems. Manage. Sci. Ser. B, Appl. Ser. **19**(8), 936–946 (1973). https://doi.org/10.1287/mnsc.19.8.936
28. G. Büyüközkan, D. Ruan, Evaluation of software development projects using a fuzzy multi-criteria decision approach. Math. Comput. Simul. **77**, 464–475 (2008). https://doi.org/10.1016/j.matcom.2007.11.015
29. J. Glasco, *Last Mile Delivery Solutions in Smart Cities and Communities*. https://hub.beesmart.city/solutions/last-mile-delivery-solutions-in-smart-cities. Last accessed 21 Sept 2018
30. G. Buyukozkan, D. Uzturk, Smart last mile delivery solution selection for cities, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, 3–5 July, 2019, London, UK, pp. 278–284
31. Bee Smart City, *MIT's Persuasive Electric Vehicle*. https://beesmart.city/solutions/mit-s-persuasive-electric-vehicle. Last accessed 26 Sept 2018

The Concept of Persvective Flexible Manufacturing System for a Collaborative Technological Cells

Vladimir Serebrenny, Dmitry Lapin, and Alisa Lapina

Abstract This paper is aimed at forming the concept of perspective manufacturing system for a collaborative technological cells. Small innovative enterprises and large production centers in the stage of modernization are typical companies in terms of the systems mentioned above. An analysis of transition requirements to Industry 4.0 under the program 'National technology initiative' and features of newly forming robotic enterprises has been performed. An overview of perspective manufacturing systems has been highlighted. The principles of structures forming and functioning mechanisms has been shown. The concept of a predictive manufacturing system has been suggested. This concept is based on the mutual integration of two entities: the heterogeneous multi-agent system with the coalitions forming mechanism and the scalable technology of end-to-end identification of complex technical systems. Primary detailing has been carried out highlighting possible advantages and disadvantages of the concept. Conclusions have been drawn about the development perspective of the studies in this direction to set future objectives.

Keywords Collaborative robotics · Identification · Manufacturing system · Multiagent system · Robotic enterprise · Robotic manufacturing · Robotic enterprise · Technological cell

1 Introduction

Basic principles and modern technologies, laying on the fourth industrial revolution, are being implemented and they have been witnessed in [1, 2]. A special place in this issue is occupied by investigating flexible manufacturing systems (FMS)

V. Serebrenny (✉) · D. Lapin · A. Lapina
Bauman Moscow State Technical University, Gospitalny Lane 10, Moscow 105005, Russia
e-mail: vsereb@bmstu.ru

D. Lapin
e-mail: lapindv@bmstu.ru

A. Lapina
e-mail: alisa.mokaeva@bmstu.ru

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8_18

which replaced traditional ones. The essence of FMS is considered as the ability to effectively produce hi-tech goods and services and rapidly react to customers' needs.

Countries all over the globe have their own ways to implement the Industry 4.0 paradigm. This study is based on principles of the Russian national technology initiative (NTI). To be more specific, this study is based on cross-market and cross-industry focus area(TechNet). Where TechNet provides technological support for the development of NTI markets and hi-tech industries by forming Factories of the Future [3].

The question of perspective flexible manufacturing systems (PFMS) implementation is one of the key technological barriers. However; considering the universal properties of the above-mentioned manufacturing systems, high efficiency and synergy of the system deployment on the specific platforms are questionable. This negotiation has been done taking into account special technical and economic circumstances.

The goal of this study is developing the specialized perspective manufacturing system for the newly forming robotic enterprises. This issue has been discussed in [4]. The following tasks have to be done:

- Estimating initial conditions of the newly forming robotic enterprises for further integration.
- Analyzing existing PFMS and highlighting their essential features and properties.
- Forming the concept of PFMS for the newly forming robotic enterprises, its structure, and operating mechanisms.
- Detailing substantive organization and observation subsystems.

2 Background

In this section, basic definitions and structures are shown to clarify the prerequisites for the concept formation. According to the NTI concept, one of the advanced manufacturing technologies are:

- Industrial sensing: It includes smart sensing and control systems embedded into manufacturing equipment, shop floor or factory.
- Industrial robotics: They are primarily flexible manufacturing cells.
- Enterprise information systems: These system include Industrial Control Systems (ICS), Manufacturing Execution Systems (MES), Enterprise Resource Planning (ERP), and Enterprise Application Software (EAS).

2.1 Factories of the Future

The overall concept of the "Factories of the Future" in accordance with the extract taken from TechNet NTI justifying materials is as follows [3].

A factory of the future is a certain kind of business process or a method of combining business processes with the following features:

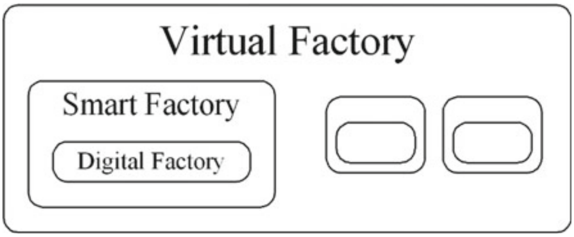


Fig. 1 Interconnection of the structures of factories of the future

- Creating digital platforms which are kind of ecosystems with advanced digital technologies. By leveraging predictive analytics and big data, the platform approach enables integration of spatially distributed designers and manufacturers. This in turn allows increasing flexibility and customization while fulfilling customer requests.
- Developing a system of digital models of both newly designed products and production processes. Digital models must have a high level of adequacy to physical products and actual processes enabling the convergence of the physical and digital worlds to generate synergy effects.
- Digitalization the entire product lifecycle from concept and design to production, use, after-sales service, and finally recycling. Later changes on a product in the lifecycle are made. Costs of production increase due to these changes. Investing more up front to get the design correct, moves changes to an earlier point in the lifecycle where the cost is less. Product quality which makes it globally competitive and appealing to consumers is also determined early in the product lifecycle, specifically during design phase i.e. before a physical representation is produced.

If taken separately, none of the advanced manufacturing technologies can provide a long-term competitive advantage in the market. Complex technological solutions are therefore needed to design and manufacture a new generation of globally competitive products in the shortest time possible. These solutions, made up of the best world-class technologies, are referred to as Digital, Smart, Virtual Factories of the Future and their interconnection in the TechNet Roadmap (Fig. 1).

2.1.1 Digital Factories

Digital Factories are complex technological solutions that ensure the design and manufacture of globally competitive next-generation products in the shortest time possible. They span the product lifecycle from the R&D and product planning stage. Product basic characteristics are set through the period from development of a digital mock-up (DMU) and digital twin up to making prototypes and small-batch production (paperless production and digital thread). A Digital Factory entails smart models of products (e.g. machines, structures, units, instruments, installations, etc.). These models are developed through the application of the new paradigm of digital design and simulation called ‘Smart Digital Twin—(Simulation & Optimization), Smart Big Data-Driven Advanced (Design & Manufacturing)’.

2.1.2 Smart Factories

Smart Factories are complex technological solutions that ensure the manufacture of globally competitive next-generation products from a workpiece to finished part in the shortest time possible. Smart manufacturing is distinctive because it provides a high level of automation and robotics. This in turn drastically reduces quality losses which have usually caused by human errors where the production is unmanned production. Smart Factories usually use digital mock-ups, digital twins and prototypes developed by Digital Factories. A Smart Factory entails manufacturing equipment which represent machine tools with computer numerical control, industrial robots, etc. Moreover; it entails industrial control systems (ICS) and manufacturing execution systems (MES).

2.1.3 Virtual Factories

Virtual Factories are complex technological solutions that ensure the design and manufacture of globally competitive next-generation products in the shortest time possible by linking Digital and/or Smart Factories into a distributed network. A Virtual Factory entails enterprise application software (EAS). This software allows developing and using virtual models of organizational, technological, logistical and other processes as a single object at the level of global supply chains (supply, production, distribution and logistics, sales, after-sales service) and/or at the level of distributed production assets.

For the newly forming robotic enterprises, it is common to use Digital Factory and Smart Factory structure features.

2.2 Newly Forming Robotics Enterprises

The newly forming robotic enterprises are part of the technological business development [3]. There are two scenarios of elaborating them: creation from scratch and modification of the existing production. Considering both the scenarios in more details, they define the key technological solutions and form requirements for effective manufacturing system.

2.2.1 Creating Enterprise from Scratch

This scenario means production deployment by small innovative enterprises [5]. The following technical and economic factors are characteristic for the development of this manufacturing:

- Small batches of the product.
- Small number of tools.
- Small areas of location.
- Rapidly changing product.

Assuming these, it is possible to conclude the following:

- Risks are related to the absence of observation because of space and resource economy.

2.2.2 Modification of the Traditional Manufacturing

This scenario means partial or full production deployment on the base of medium-sized innovative enterprises and large production centers [6]. The following technical and economic factors are characteristic for the development of this manufacturing:

- Large batches of the product.
- Large number of tools.
- Large areas of location.
- Temperately changing product.

Assuming these, it is possible to conclude the following:

- Risks are related to the absence of organization because of difficulties of integration in the existing system.

2.3 Critical Technologies

Each type of enterprise has the following critical technological solutions

- Collaborative robotics [7].
- Big data [8].
- Sets of sensors [9].

The high intelligence level of the modern tools also impacts the concept when discussing the effective use of intelligent elements of these solutions at all levels of management.

2.4 Technical and Economic Factors

In both cases, minimizing risks and achieving high efficiency can be reached by implementing lean principles. Proven principles of lean—such as reducing waste in the form of machine breakdowns or non-value—adding activities—remains fundamental. At the same time, advancements in data collection, sensors, robotics and automation, new technologies and increased computing power enable the advanced analytics and give the established methods a new edge [10].

In this regard, the most rational way is to establish the concept aimed at solving the main problems of the above-described enterprises. These problems include formation of the flexible manufacturing organizing system, and introduction of universal scalable observation system. Moreover; these problems include the result of interaction between a control system which is based on reconfigurable manufacturing systems approach and the development perspectives to adaptive manufacturing system [11, 12].

3 Analysis

To enable efficient use of the data available in Industry 4.0 oriented processes, it becomes necessary to adapt the control architecture in order to make it flexible, reactive and adaptable enough to reach the objectives previously described. For the last 20 years, Holonic Control Architectures (HCA) has been widely studied and developed. The use at the industrial level is starting to spread due to their effectiveness. A Holonic Control Architecture (HCA) is composed of holons, called holarchy. A holon is a communicating decisional entity (with inputs and outputs) composed of a set of sub-level holons and at the same time, part of a wider organization composed of higher-level holons (recursive, called the Janus effect [13]). It is important to note that a holon is also composed of a physical part associated with a digital one (that can be modeled as a digital agent, avatar, digital twin) and finally, holons are able to decide according to a certain degree of autonomy [14].

Dynamic HCA are interesting because they integrate an optimal scheduling module used in the normal state, coupled with reactive abilities, executed when a disturbance occurs. When it happens, the HCA may modify its own organization to minimize the impact of this disturbance. Such architectures guarantee that performances of the manufacturing system are optimal in the normal state, but not always in degraded mode.

3.1 ORCA-FMS

In the manufacturing domain, ORCA—dynamic and heterogeneous hybrid Architecture for Optimized and Reactive Control—was one of the first dynamic HCA formalized in literature [15]. ORCA is divided into three layers: the physical system (PS) layer, the local control (LC) layer, and the global control (GC) layer. The GC has a global view of the system and is composed of one global optimizer. Its role is to guarantee good performances based on the PS's initial state. The LC is composed of many local optimizers, which have a local view of the system only. Their goal is to react to unexpected events that occur on the PS by making rapid online decisions. Thus; they provide a feasible solution which is suitable for the current system state. Each local optimizer in LC is associated with a physical part in PS. Both the physical part and the local optimizer constitute one entity.

3.2 ADACOR

ADACOR (ADAptive holonic COntrol aRchitecture) is a holonic reference for the distributed manufacturing system [16]. ADACOR is a decentralized control architecture but it also considers centralization in order to tend to global optimization of the system. Holons belong to the following classes: Product Holons (ProdH), Task Holons (TH), Operation Holons (OpH) and Supervisor Holons (SupH).

An evolution of ADACOR mechanism has also been presented as ADACOR2 [17]. The objective is to let the system evolve dynamically through configurations discovered online, not only between a stationary and one transient state. The rest of the architecture is nevertheless quite similar to ADACOR.

3.3 POLLUX

The last architecture in date is denoted as POLLUX [18]. The main novelty is centralized in the adaptation mechanism of the architecture. It uses governance parameters to enlarge or constraint the behavior of the low-level holons regarding the disturbances observed by the higher level. Therefore; the idea is to find the "best" architecture that suits detected disturbances. It is a Hybrid Control Architecture (HCA) presented as a reference control system that supports the switching process of the control system between hierarchical and heterarchical architectures.

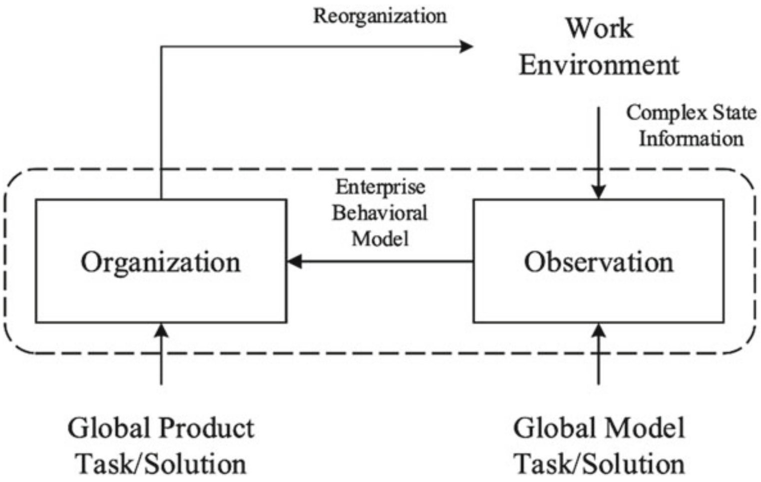

Reorganization
Work Environment
Complex State Information
Enterprise Behavioral Model
Organization
Observation
Global Product Task/Solution
Global Model Task/Solution

Fig. 2 Subsystems implementation and their interaction mechanisms

4 Concept

In comparison with the above-mentioned PMS, the concept of the manufacturing system is proposed.

4.1 Description

The system is based on dynamic organizing and observation subsystems. Subsystems implementation and their interaction mechanisms are shown on Fig. 2.

This effect is achieved due to the implementation features of the organizing and monitoring subsystems.

Organizing subsystem is based on the multi-agent systems approach and on observing subsystem which is end-to-end structural and parametric identification tool.

4.2 Structure and Mechanism

The overall structure of the proposed production system is based on two subsystems that carry out both independent functioning and mutual influence. The total effect of adaptive management is due to the effective decomposition of the global task, as well as through the continuous exchange of information at all levels of control (Fig. 3).

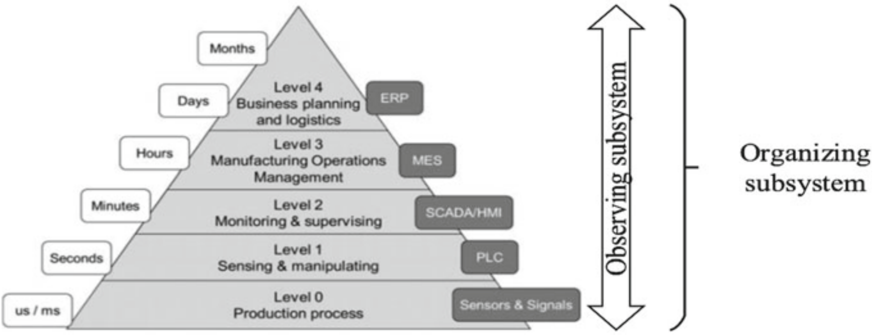


Fig. 3 Subsystems implementation and their interaction mechanisms

Subsystems exchange mechanism is based on complex behavioral model of the enterprise (Fig. 5). The main object in the behavioral setting is the “behavior” which is a set of all signals compatible with the system. An important feature of this approach is that it does not distinguish between the priority of input and output variables. Apart from putting system theory and control on a rigorous basis, the behavioral approach has unified the existing approaches. Moreover; it has brought new results on controllability for nD systems, control via interconnection, and system identification [19].

5 Detailing

This section describes more detailed consideration of the organizing and monitoring subsystems structures.

5.1 Multi-agent System

The system used to organize manufacturing tools is a multi-agent group control system that uses a dynamic mechanism of homogeneous and/or heterogeneous coalitions formation [20, 21]. The implementation of such a mechanism brings the system closer to the hybrid control architecture. However, system formation rules are associated with the identification tool result in interpretation in the real-time.

Analysis of the modern multi-agent architectures showed that of greatest interest is the coalition model of the system, the essence of which lies in the subgroups formation of agents. Each group may be considered as a separate self-sufficient system, and as part of the global system. For example, homogenous coalitions are good in a quick readjusting of the production line, heterogeneous—with the selected manufacturing process (Fig. 4).

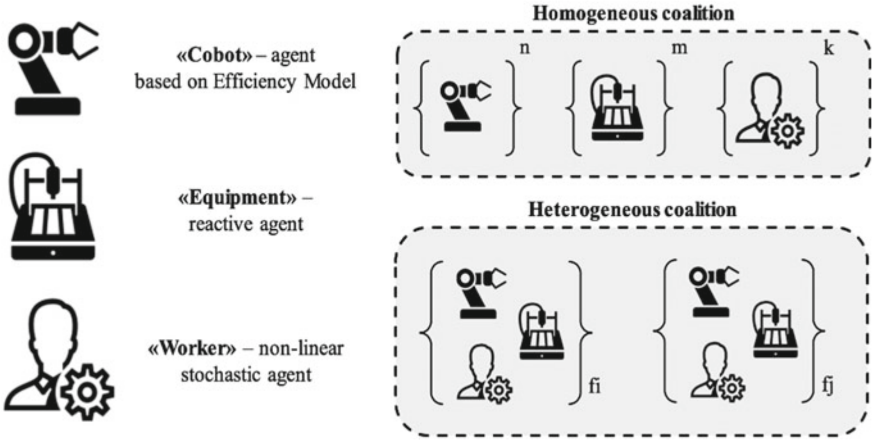

«Cobot» – agent based on Efficiency Model
«Equipment» – reactive agent
«Worker» – non-linear stochastic agent
Homogeneous coalition
n
m
k
Heterogeneous coalition
fi
fj

Fig. 4 Subsystems implementation and their interaction mechanisms

5.2 *Identification Tools*

The system used to observe manufacturing tools is an end-to-end structural and parametric identification tool. The math part is based on wavelet transforms theory.

Wavelet transforms of the input and output parameters are proposed for organizing tools for end-to-end processes identification (Fig. 5) [22, 23].

Using the real-time representation of the system in the combined time-frequency domain, it is possible to critically evaluate linear and non-linear processes in the system as well as obtaining interpretable information about the real generalized structure and parameters of the system.

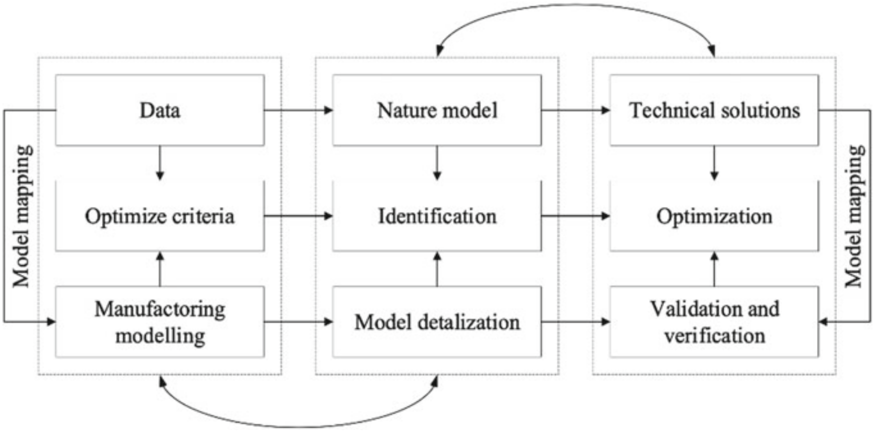


Fig. 5 Subsystems implementation and their interaction mechanisms

6 Conclusion and Future Work

After examining results of this study, it is possible to conclude the following points:

- The concept of perspective manufacturing system for the newly forming robotic enterprises meet the modern requirements and technological barriers of the National development program.
- The key difference with the previously developed PMS is the desire for local efficiency in a particular type of robotic enterprises.
- The use of combined advantages of the organization and observation subsystems allows achieving a synergistic effect.

Future work will consist of system specification for further software and hardware implementation.

References

1. Kagermann et al. Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0. (2018, December, 23). Retrieved from https://www.bmbf.de/files/Umsetzungsempfehlungen_Industrie4_0.pdf
2. V.B. Tarassov, Enterprise total agentification as a way to industry 4.0: forming artificial societies via goal-resource networks, in *International Conference on Intelligent Information Technologies for Industry* (Springer, Cham, 2018, September), pp. 26–40
3. TechNet booklet (2018, March, 7). Retrieved from http://assets.fea.ru/uploads/fea/nti/docs/TechNET_booklet_ENG_web%20version_2018_02_07.pdf
4. V. Serebrenny, D. Lapin, A. Mokaeva, The concept of flexible manufacturing system for a newly forming robotic enterprises, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2019*, 3–5 July, 2019, London, UK, pp. 267–271
5. Z.M. Bi, Y. Liu, B. Baumgartner, E. Culver, J.N. Sorokin, A. Peters, S. O'Shaughnessey et al., Reusing industrial robots to achieve sustainability in small and medium-sized enterprises (SMEs). Ind. Robot Int. J. **42**(3), 264–273 (2015)
6. M. Hedelind, M. Jackson, How to improve the use of industrial robots in lean manufacturing systems. J. Manuf. Technol. Manag. **22**(7), 891–905 (2011)
7. A. Vysocky, P. Novak, Human-Robot collaboration in industry. MM Sci. J. **9**(2), 903–906 (2016)
8. J. Lee, B. Bagheri, H.A. Kao, Recent advances and trends of cyber-physical systems and big data analytics in industrial informatics, in *International Proceeding of Int Conference on Industrial Informatics (INDIN)* (2014, July), pp. 1–6
9. D. Strang, R. Anderl, Assembly process driven component data model in cyber-physical production systems, in *Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering and Computer Science 2014*, 22–24 October, 2014, San Francisco, USA, pp. 22–24
10. A. Behrendt, N. Müller, P. Odenwälder, C. Schmitz, Industry 4.0 demystified—lean's next level (2017). Retrieved March, 3, 2017
11. Y. Koren, U. Heisel, F. Jovane, T. Moriwaki, G. Pritschow, G. Ulsoy, H. Van Brussel, Reconfigurable manufacturing systems. CIRP Annals **48**(2), 527–540 (1999)
12. N. Keddis, G. Kainz, C. Buckl, A. Knoll, Towards adaptable manufacturing systems, in *2013 IEEE International Conference on Industrial Technology (ICIT)* (IEEE, 2013, February), pp. 1410–1415

13. A. Koestler, *Der Mensch-Irrläufer der Evolution* (Scherz, Bern München, 1978)
14. R.F. Babiceanu, F.F. Chen, Development and applications of holonic manufacturing systems: a survey. J. Intell. Manuf. **17**(1), 111–131 (2006)
15. C. Pach, T. Berger, T. Bonte, D. Trentesaux, ORCA-FMS: a dynamic architecture for the optimized and reactive control of flexible manufacturing scheduling. Comput. Ind. **65**(4), 706–720 (2014)
16. P. Leitão, F. Restivo, ADACOR: a holonic architecture for agile and adaptive manufacturing control. Comput. Ind. **57**(2), 121–130 (2006)
17. J. Barbosa, P. Leitão, E. Adam, D. Trentesaux, Dynamic self-organization in holonic multi-agent manufacturing systems: the ADACOR evolution. Comput. Ind. **66**, 99–111 (2015)
18. J.F. Jimenez, A. Bekrar, G. Zambrano-Rey, D. Trentesaux, P. Leitão, Pollux: a dynamic hybrid control architecture for flexible job shop systems. Int. J. Prod. Res. **55**(15), 4229–4247 (2017)
19. I. Markovsky, J.C. Willems, S. Van Huffel, B. De Moor, *Exact and Approximate Modeling of Linear Systems: A Behavioral Approach*, vol. 11 (SIAM, 2006)
20. S. Vorotnikov, K. Ermishin, A. Nazarova, A. Yuschenko, Multi-agent Robotic Systems in Collaborative Robotics, in *International Conference on Interactive Collaborative Robotics* (Springer, Cham, 2018, September), pp. 270–279
21. M. Pechoucek, V. Marik, O. Stepankova, Coalition formation in manufacturing multi-agent systems, in *Proceedings 11th International Workshop on Database and Expert Systems Applications* (IEEE, 2000), pp. 241–246
22. N. Bakhtadze, E. Sakrutina, Wavelet-based identification and control of variable structure systems, in *2016 International Siberian Conference on Control and Communications (SIBCON)* (IEEE, 2016, May), pp. 1–6
23. N. Bakhtadze, E. Sakrutina, Applying the multi-scale wavelet-transform to the identification of non-linear time-varying plants. IFAC-Papers OnLine **49**(12), 1927–1932 (2016)

Index

S.-I. Ao et al. (eds.), *Transactions on Engineering Technologies*,
https://doi.org/10.1007/978-981-15-8273-8